Title
On the Design and Analysis of Operator-Splitting Schemes

Permalink
https://escholarship.org/uc/item/41c3v36v

Author
Davis, Damek

Publication Date
2015

Peer reviewed|Thesis/dissertation
On the Design and Analysis of Operator-Splitting Schemes

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

by

Damek Shea Davis

2015
This thesis is concerned with the design and analysis of algorithms that solve nonsmooth convex optimization problems in (possibly infinite dimensional) Hilbert spaces. There are many algorithms available to solve such problems, but the methods detailed in this thesis are particularly well-suited for solving complicated problems that are built from many simpler pieces. There are a wealth of applications for which such structure is present, and this has driven the recent resurgence of interest in so-called Operator-Splitting methods; these splitting methods completely disentangle complex problem structure and give rise to algorithms that repeatedly solve a series of simpler subproblems sequentially or in parallel. These algorithms are easy to implement, they have low per-iteration cost, and in practice, they are observed to quickly converge to solutions of modest accuracy. These qualities make splitting methods attractive options for solving large-scale problems in machine learning and signal processing.

Although splitting algorithms are known to converge, a general theoretical analysis of their convergence rates has remained elusive since their inception nearly in the 1950s. Furthermore, since 2000, no new splitting algorithms have been developed that do not reduce to one of three general methods (for solving general monotone inclusion problems). The purpose of this thesis is to address these theoretical challenges by deriving sharp convergence rates of existing splitting algorithms and developing a new splitting method that
does not appear to reduce to any existing method. The analysis presented is particularly simple, and applies in many cases.
The dissertation of Damek Shea Davis is approved.

Lieven Vandenberghe

Stanley Osher

Stefano Soatto

Wotao Yin, Committee Chair

University of California, Los Angeles
2015
To Nicole . . .

For never insisting
that I be anything other
than who I am
# Table of Contents

1 Introduction .......................................................... 1

2 Convergence rate analysis of several splitting schemes ............ 4
  2.1 Introduction .......................................................... 4
     2.1.1 Goals, challenges, and approaches .......................... 5
     2.1.2 Notation ............................................................ 8
     2.1.3 Assumptions ...................................................... 9
     2.1.4 The Algorithms .................................................. 9
     2.1.5 Basic properties of averaged operators ....................... 12
  2.2 Summable sequence convergence lemma ............................. 13
  2.3 Iterative fixed-point residual analysis .............................. 16
     2.3.1 o(1/(k + 1)) FPR of averaged operators ..................... 17
     2.3.2 o(1/(k + 1)) FPR of relaxed PRS ............................ 20
     2.3.3 o(1/(k + 1)^2) FPR of FBS and PPA ......................... 21
     2.3.4 o(1/(k + 1)^2) FPR of one dimensional DRS .................. 24
     2.3.5 O(1/\Lambda_k^2) ergodic FPR of Fejér monotone sequences .... 25
  2.4 Subgradients and fundamental inequalities .......................... 26
     2.4.1 A subgradient representation of relaxed PRS ................. 27
     2.4.2 Optimality conditions of relaxed PRS ....................... 28
     2.4.3 Fundamental inequalities ....................................... 29
  2.5 Objective convergence rates ........................................ 30
     2.5.1 Ergodic convergence rates ..................................... 31
2.5.2 Nonergodic convergence rates ........................................... 33
2.6 Optimal FPR rate and arbitrarily slow convergence ................. 37
  2.6.1 Optimal FPR rates ...................................................... 37
  2.6.2 Arbitrarily slow convergence ....................................... 40
2.7 Optimal objective rates ..................................................... 42
  2.7.1 Ergodic convergence of feasibility problems ....................... 42
  2.7.2 Ergodic convergence of minimization problems .................... 42
  2.7.3 Optimal nonergodic objective rates ................................. 45
  2.7.4 Optimal objective and FPR rates with Lipschitz derivative ..... 51
2.8 From relaxed PRS to relaxed ADMM ..................................... 52
  2.8.1 Dual feasibility convergence rates .................................. 54
  2.8.2 Converting dual inequalities to primal inequalities ............... 55
  2.8.3 Converting dual convergence rates to primal convergence rates .. 56
2.9 Examples ........................................................................... 58
  2.9.1 Feasibility problems ...................................................... 58
  2.9.2 Parallelized model fitting and classification ......................... 59
  2.9.3 Distributed ADMM ....................................................... 62
2.10 Conclusion ................................................................. 64

3 Faster convergence rates of relaxed Peaceman-Rachford and ADMM
under regularity assumptions .................................................. 66
  3.1 Introduction ............................................................... 66
    3.1.1 Goals, challenges, and approaches ................................ 67
    3.1.2 Notation ............................................................... 70
    3.1.3 Assumptions ......................................................... 71
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.4</td>
<td>The Algorithms</td>
<td>72</td>
</tr>
<tr>
<td>3.1.5</td>
<td>Practical implications: a comparison with FBS</td>
<td>73</td>
</tr>
<tr>
<td>3.1.6</td>
<td>Basic properties of proximal operators</td>
<td>74</td>
</tr>
<tr>
<td>3.1.7</td>
<td>Convergence rates of summable sequences</td>
<td>74</td>
</tr>
<tr>
<td>3.1.8</td>
<td>Convergence of the fixed-point residual (FPR)</td>
<td>76</td>
</tr>
<tr>
<td>3.1.9</td>
<td>Subgradients</td>
<td>77</td>
</tr>
<tr>
<td>3.1.10</td>
<td>Fundamental inequalities</td>
<td>78</td>
</tr>
<tr>
<td>3.2</td>
<td>Strong convexity</td>
<td>80</td>
</tr>
<tr>
<td>3.3</td>
<td>Lipschitz derivatives</td>
<td>82</td>
</tr>
<tr>
<td>3.3.1</td>
<td>The general case: best iterate convergence rate</td>
<td>83</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Constant relaxation and better rates</td>
<td>87</td>
</tr>
<tr>
<td>3.4</td>
<td>Linear convergence</td>
<td>94</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Solely regular ( f ) or ( g )</td>
<td>96</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Complementary regularity of ( f ) and ( g )</td>
<td>98</td>
</tr>
<tr>
<td>3.5</td>
<td>Feasibility Problems with regularity</td>
<td>99</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Multiple sets</td>
<td>108</td>
</tr>
<tr>
<td>3.6</td>
<td>From relaxed PRS to ADMM</td>
<td>111</td>
</tr>
<tr>
<td>3.6.1</td>
<td>Converting dual inequalities to primal inequalities</td>
<td>115</td>
</tr>
<tr>
<td>3.6.2</td>
<td>Converting dual convergence rates to primal convergence rates</td>
<td>116</td>
</tr>
<tr>
<td>3.7</td>
<td>Examples</td>
<td>121</td>
</tr>
<tr>
<td>3.7.1</td>
<td>Feasibility problems</td>
<td>122</td>
</tr>
<tr>
<td>3.7.2</td>
<td>Parallelized model fitting and classification</td>
<td>125</td>
</tr>
<tr>
<td>3.7.3</td>
<td>Linear and semidefinite programming</td>
<td>128</td>
</tr>
<tr>
<td>3.8</td>
<td>Conclusion</td>
<td>131</td>
</tr>
</tbody>
</table>
4 Convergence Rate Analysis of Primal-Dual Splitting Schemes

4.1 Introduction
  4.1.1 Goals, challenges, and approaches
  4.1.2 Definitions, notation and some facts
  4.1.3 Assumptions
  4.1.4 Basic properties of metrics
  4.1.5 Basic properties of resolvents and averaged operators
  4.1.6 Variable metrics

4.2 The unifying scheme
  4.2.1 Problem and algorithm
  4.2.2 Examples of the unifying scheme
  4.2.3 The fundamental inequality

4.3 Ergodic convergence

4.4 Nonergodic convergence

4.5 Applications
  4.5.1 Primal-dual gap functions
  4.5.2 Two algorithm classes
  4.5.3 Second metric class
  4.5.4 New and old convergence rates

4.6 Conclusion

5 Convergence Rate Analysis of the Forward-Douglas-Rachford Splitting Scheme

5.1 Introduction
  5.1.1 Goals, challenges, and approaches
List of Figures

2.1 A single relaxed PRS iteration, from \( z \) to \((T_{PRS})_\lambda(z)\). .......................... 27

2.2 Illustration of Example 2.6.1. Each pair of lines represents a 2-dimensional component of \( U \cup V \). The angles \( \theta_k \) are converging to 0. .......................... 38

2.3 Example 2.7.1 of PRS. \( z^k \) hops between \((1, 1)\) and \((-1, -1)\) while the ergodic iterates \( \overline{x}_g^k \) and \( \overline{x}_f^k \) (dots of decreasing size) approach \( x^* \). ............... 43

3.1 A single relaxed PRS iteration starting from \( z \). ........................................... 77

5.1 A single FDRS iteration, from \( z \) to \((T_{FDRS})_\lambda(z)\) (see Lemma 5.2.1). Both occurrences of \( \tilde{\nabla}_V(x_h) \) represent the same subgradient \( (1/\gamma)P_{V^\perp}z = (1/\gamma)(z - x_h) \in V^\perp \). .................................................. 192

6.1 The mapping \( T : z^k \mapsto z^{k+1} := Tz^k \). The vectors \( u_B^k \in Bx_B^k \) and \( u_A^k \in Ax_A^k \) are defined in Lemma 6.3.1. ................................. 233

6.2 Image inpainting with texture completion .................................................. 246
  a Original image ............................................. 246
  b Occluded image 1 ........................................... 246
  c Occluded image 2 ........................................... 246
  d Recovered image 1 ........................................... 246
  e Recovered image 2 ........................................... 246

6.3 Movie recommendations with matrix completion ........................................ 248
  a Fixed-point residual at iteration \( k \). ................................. 248
  b Rank at iteration \( k \) ........................................... 248
  c Root mean square error (Equation (6.5.1)) at iteration \( k \) .................. 248

6.4 Classification with Support Vector Machines ........................................... 251
<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Fixed-point residual with and without line search (LS).</td>
<td>251</td>
</tr>
<tr>
<td>b</td>
<td>Objective value with and without line search (LS).</td>
<td>251</td>
</tr>
<tr>
<td>c</td>
<td>Comparison of ergodic and nonergodic iterates.</td>
<td>251</td>
</tr>
<tr>
<td></td>
<td><strong>6.5 Portfolio optimization</strong></td>
<td>253</td>
</tr>
<tr>
<td>a</td>
<td>Distance to solution using accelerated and non accelerated methods.</td>
<td>253</td>
</tr>
<tr>
<td>b</td>
<td>Objective value of accelerated and non accelerated methods. Note</td>
<td></td>
</tr>
<tr>
<td></td>
<td>that the blue and green curves perfectly overlap.</td>
<td>253</td>
</tr>
</tbody>
</table>
2.1 Overview of the main identities used throughout the chapter. The letter $s$ denotes a superscript (e.g. $s = k$ or $s = *$). The vector $z^s \in \mathcal{H}$ is an arbitrary input point. See Lemma 2.4.1 for a proof. 28

3.1 Overview of the main subgradient identities used throughout Section 3.6. The letter $s$ denotes a superscript (e.g. $s = k$ or $s = *$). See Chapter 2 for a proof. 114

4.1 This table lists the original appearance of the algorithms constructed from pairing the metrics in Section 4.5.2 with the PPA, FBS, PRS, and FBF algorithms applied to Problem 4.5.1. See Propositions 4.5.1 and 4.5.2 for the definitions of the “level.” 178

6.1 Classification accuracy for different choices of $C$ and $\sigma$ in the SVM model. 250
ACKNOWLEDGMENTS

Over the past 5 years, my life has benefited from the continued support and guidance of my friends, collaborators, and academic advisors.

Since the beginning of my graduate career at UCLA, Farzin Barekat, Bill Chen, and Chris Scaduto have been a continual source of inspiration, support, and friendship for me.

My advisor Professor Stefano Soatto brought me into his lab when I knew absolutely nothing about computer vision and machine learning. I am not sure where I would be without his support during my critical transition to applied mathematics.

My advisor Professor Wotao Yin serendipitously arrived at UCLA while I was beginning to transition to optimization. I fear that I might have fallen through the cracks without his guidance, excitement, and collaboration.

Throughout my time at UCLA, I’ve also had the pleasure of collaborating and interacting with Jonathan Balzer, Virginia Estellers, and Alex Sadovsky who helped me develop as a researcher.

A long list of people have affected my life in a very positive way: Jingming Dong, Brent Edmunds, Georgios Georgiadis, Josh Hernandez, Vasily Karasev, Nikos Karianakis, Daniel O’Connor, Zhimin Peng, Wei Shi, Brian Taylor, Konstantine Tsotsos, Tianyu Wu, and Ming Yan.

I also want to thank the other members of my committee, Professors Stanley Osher and Lieven Vandenberghe for their continued support and confidence.

All of the chapters in this manuscript have been submitted for publication. I will now briefly describe where they are currently available and the coauthor contributions.

Wotao Yin was a coauthor of Chapters 2, 3, and 6 and helped direct the research for these projects.

Chapters 4 and 5 are versions of [55] and [56] respectively.

Finally, I would like to acknowledge the support of NSF GRFP grant DGE-0707424.
VITA

2009  Phi Beta Kappa (elected junior year).

2010  B.S. Honors in Mathematics *summa cum laude*, University of California, Irvine.

2010-2015  National Science Foundation Graduate Research Fellowship Program (GRFP) in Mathematics.

2013-2014  Research Assistant, Computer Science Department, UCLA.

2015  Research Assistant, Mathematics Department, UCLA.
CHAPTER 1

Introduction

In broad terms, optimization is the study of minima and maxima of functions. Since the introduction of Fermat’s principle, the theory of optimization has undergone a tremendous development. While continuous optimization was classically confined to smooth functions over nice domains, today it is possible to work with nonsmooth functions in infinite dimensional spaces over general domains. It is precisely in this nonsmoothness that modern applications of optimization in statistical and machine learning find their power. However, nonsmooth optimization is certainly more complex than its smooth counterpart, which makes the development and analysis of general and efficient numerical schemes that much more difficult.

This thesis is concerned with the design and analysis of algorithms that solve nonsmooth convex optimization problems in (possibly infinite dimensional) Hilbert spaces. There are many algorithms available to solve such problems, but the methods detailed in this thesis are particularly well-suited for solving complicated problems that are built from many simpler pieces. There are a wealth of applications for which such structure is present, and this has driven the recent resurgence of interest in so-called Operator-Splitting methods; these splitting methods completely disentangle complex problem structure and give rise to algorithms that repeatedly solve a series of simpler subproblems sequentially or in parallel. These algorithms are easy to implement, they have low per-iteration cost, and in practice, they are observed to quickly converge to solutions of modest accuracy. These qualities make splitting methods attractive options for solving large-scale problems in machine learning and signal processing.
Although splitting algorithms are known to converge, a general theoretical analysis of their convergence rates has remained elusive since their inception in the 1950s. Furthermore, since 2000, no new splitting algorithms have been developed that do not reduce to one of three general methods (for solving general monotone inclusion problems). The purpose of this thesis is to address these theoretical challenges by deriving sharp convergence rates of existing splitting algorithms and developing a new splitting method that does not appear to reduce to any existing method. The analysis presented is particularly simple and applies in many cases.

Chapter 2 is the most foundational of all of the chapters. It contains the convergence rate analysis of the Douglas-Rachford and ADMM splitting schemes, but along the way, it develops several general tools that will be applied repeatedly in the subsequent chapters.

Chapter 3 investigates whether the sharp convergence rates of Chapter 2 can be improved when more assumptions are imposed on the problem, such as differentiability and strong convexity.

Chapter 4 investigates the convergence rate of a particular class of splitting schemes called *primal-dual*, which are far more powerful than the methods investigated in Chapter 2. The results of this chapter rely more heavily on the techniques of monotone operator than the previous ones do.

Chapter 5 investigates the convergence rate of the forward-Douglas-Rachford splitting scheme and presents a rather compact analysis that further develops the techniques used in Chapters 2 and 3.

Chapter 6 concludes this thesis with the development of a new splitting method. It contains, as a special case, many of the algorithms investigated in the previous chapters. Several experiments, accelerations, and enhancements, as well as a detailed convergence rate analysis are presented.

The theoretical tools used throughout this thesis are found in convex analysis and monotone operator theory—fields which have been rapidly developing since the 1950s.
Due to the complexity of these fields, we assume a fairly high level of familiarity with these tools, and the interested reader is recommended to follow the development of the subject contained in the recent monograph [11], which we use as our primary reference. However, to make life easier on the reader, each chapter is written to be self-contained and independent of the other chapters, insofar as it is possible. Thus, we always include notation and standard definitions at the start of each chapter and a brief introduction to remind the reader why the chapter should be interesting to them.
CHAPTER 2

Convergence rate analysis of several splitting schemes

2.1 Introduction

Operator-splitting and alternating-direction methods have a long history, and they have been, and still are, some of the most useful methods in scientific computing. These algorithms solve problems composed of several competing structures, such as finding a point in the intersection of two sets, minimizing the sum of two functions, and, more generally, finding a zero of the sum of two monotone operators. They give rise to algorithms that are simple to implement and converge quickly in practice. Since the 1950s, operator-splitting methods have been applied to solving partial differential equations (PDEs) and feasibility problems. Recently, certain operator-splitting methods such as ADMM (for alternating direction methods of multipliers) [64, 65] and Split Bregman [68] have found new applications in (PDE and non-PDE related) image processing, statistical and machine learning, compressive sensing, matrix completion, finance, and control. They have also been extended to handle distributed and decentralized optimization (see [26, 110, 114]).

In convex optimization, operator-splitting methods split constraint sets and objective functions into subproblems that are easier to solve than the original problem. Throughout this chapter, we will consider two prototype optimization problems: We analyze the unconstrained problem

$$\min_{x \in \mathcal{H}} f(x) + g(x)$$  \hspace{1cm} (2.1.1)
where $\mathcal{H}$ is a Hilbert space. In addition, we analyze the linearly constrained variant

$$\begin{align*}
\text{minimize} & \quad f(x) + g(y) \\
\text{subject to} & \quad Ax + By = b
\end{align*}$$

(2.1.2)

where $\mathcal{H}_1, \mathcal{H}_2,$ and $\mathcal{G}$ are Hilbert spaces, the vector $b$ is an element of $\mathcal{G}$, and $A : \mathcal{H}_1 \to \mathcal{G}$ and $B : \mathcal{H}_2 \to \mathcal{G}$ are bounded linear operators. Our working assumption throughout the chapter is that the subproblems involving $f$ and $g$ separately are much simpler to solve than the joint minimization problem.

Problem (2.1.1) is often used to model tasks in signal recovery that enforce prior knowledge of the form of the solution, such as sparsity, low rank, and smoothness [46]. The knowledge-enforcing function, also known as the regularizer, often has properties that make it difficult to jointly optimize with the remaining parts of the problem. Therefore, operator-splitting methods become the natural choice.

Problem (2.1.2) is often used to model tasks in machine learning, image processing, and distributed optimization. For special choices of $A$ and $B$, operator-splitting schemes naturally give rise to algorithms with parallel or distributed implementations [18, 26]. Because of the flexibility of operator-splitting algorithms, they have become a standard tool that addresses the emerging need for computational approaches to analyze a massive amount of data in a fast, parallel, distributed, or even real-time manner.

### 2.1.1 Goals, challenges, and approaches

This work seeks to improve the theoretical understanding of the most well-known operator-splitting algorithms including Peaceman-Rachford splitting (PRS), Douglas-Rachford splitting (DRS), the alternating direction method of multipliers (ADMM), as well as their relaxed versions. (The proximal point algorithm (PPA) and forward-backward splitting (FBS) are covered in some limited aspects too.) When applied to convex optimization problems, they are known to converge under rather general conditions. However, their objective error convergence rates are largely unknown with only a few exceptions.
Among the convergence rates known in the literature, many are given in terms of quantities that do not immediately imply the objective error rates, such as the fixed-point residual (FPR) \(^{[72, 51, 82]}\) (the squared distance between two consecutive iterates). Furthermore, almost all (with the exception of \([14]\) and a part of \([108]\)) analyze the objective error or variational inequalities evaluated at the time-averaged, or ergodic, iterate, rather than the last, or nonergodic, iterate generated by the algorithm \([36, 21, 91, 89, 90]\). In applications, the nonergodic iterate tends to share structural properties with the solution of the problem, such as sparsity in \(\ell_1\) minimization or low-rankness in nuclear norm minimization. In contrast, the ergodic iterates tend to “average out” these properties in the sense that the average of many sparse vectors can be dense. Thus, part of the purpose of this work is to illustrate the theoretical differences between these two iterates and to justify the use of nonergodic iterates in practice.

Unlike the well-developed complexity estimates for (sub)gradient methods \([93]\), there are no known lower complexity results for most strictly primal or strictly dual splitting algorithms. (As an exception, lower complexity results are known for the primal-dual case \([41]\).) In particular, the classical complexity analysis given in \([93, 94]\) does not apply to the algorithms addressed in this chapter. This chapter attempts to close this gap. The convergence rates in terms of FPR and objective errors are derived for operator-splitting algorithms applied to Problem (2.1.1). Convergence rates for constraint violations, the primal objective error, and the dual objective error are derived for ADMM, which applies to Problem (2.1.2). Some of the derived rates are convenient to use, for example, to determine how many iterations are needed to reach a certain accuracy, to decide when to stop an algorithm, and to compare an algorithm to others in terms of their worst-case complexities.

The techniques we develop in this chapter are quite different from those used in classical optimization convergence analysis, mainly because splitting algorithms are driven by fixed-point operators instead of driven by minimizing objectives. (An exception is FBS, which is driven by both.) Splitting algorithms are fixed-point iterations derived from
certain optimality conditions, and they converge due to the contraction of the fixed-point operators. Some of them do not even reduce objectives monotonically. Thus, objective convergence is a consequence of operator convergence rather than the cause of it. Therefore, we first perform an operator-theoretic analysis and then, based on these results, derive optimization related rates.

We now describe our contributions and our techniques as follows:

- We show that the FPR of the fixed-point iterations of nonexpansive operators converge with rate $o(1/(k+1))$ (Theorem 2.3.1). This rate is optimal and improves on the known big-O rate [51, 82]. For the special cases of FBS applied to Problem (2.1.1) and one-dimensional DRS, we improve this rate to $o(1/(k+1)^2)$. In addition, we provide examples (Section 2.6) to show that all of these rates are tight. Specifically, for each rate $o(1/(k+1)^p)$, we give an example with rate $\Omega(1/(k+1)^{p+\varepsilon})$ for any $\varepsilon > 0$. A detailed list of our contributions and a comparison with existing results appear in Section 2.6.1.1. The analysis is based on establishing summable and, in many cases, monotonic sequences, whose convergence rates are summarized in Lemma 2.2.1.

- We demonstrate that even when the DRS algorithm converges in norm to a solution, it may do so arbitrarily slowly (Theorem 2.6.2).

- We give the objective convergence rates of the relaxed PRS algorithm and show that it is, in the worst case, nearly as slow as the subgradient method (Theorems 2.5.2 and 2.7.2), yet nearly as fast as PPA in the ergodic sense (Theorems 2.5.1 and 2.7.3). The rates are obtained by relating the objective error to the aforementioned FPR rates through a fundamental inequality (Proposition 2.4.1). Note that we prove ergodic and nonergodic convergence rates and show that these rates are sharp through several examples (Section 2.7).

- We give the convergence rates of the primal objective and feasibility of the current iterates generated by ADMM. Our analysis follows by a simple application of the
2.1.2 Notation

In what follows, \( \mathcal{H}, \mathcal{H}_1, \mathcal{H}_2, \mathcal{G} \) denote (possibly infinite dimensional) Hilbert spaces. In fixed-point iterations, \((\lambda_j)_{j \geq 0} \subset \mathbb{R}_+\) will denote a sequence of relaxation parameters and 

\[
\Lambda_k := \sum_{i=0}^{k} \lambda_i
\]

is its \(k\)th partial sum. To ease notational memory, the reader may assume that \(\lambda_k \equiv (1/2)\) and \(\Lambda_k = (k + 1)/2\) in the DRS algorithm or that \(\lambda_k \equiv 1\) and \(\Lambda_k = (k + 1)\) in the PRS algorithm. Given the sequence \((x^j)_{j \geq 0} \subset \mathcal{H}\), we let \(\bar{x}^k = (1/\Lambda_k) \sum_{i=0}^{k} \lambda_i x^i\) denote its \(k\)th average with respect to the sequence \((\lambda_j)_{j \geq 0}\).

We call a convergence result \textit{ergodic} if it is in terms of the sequence \((\bar{x}^j)_{j \geq 0}\), and \textit{nonergodic} if it is in terms of \((x^j)_{j \geq 0}\).

Given a closed, proper, convex function \(f : \mathcal{H} \to (-\infty, \infty]\), \(\partial f(x)\) denotes its subdifferential at \(x\) and

\[
\tilde{\nabla} f(x) \in \partial f(x), 
\]

(2.1.3)

denotes a subgradient, and the actual choice of the subgradient \(\tilde{\nabla} f(x)\) will always be clear from the context. (This notation was used in [16, Eq. (1.10)].)

The convex conjugate of a proper, closed, and convex function \(f\) is

\[
f^*(y) := \sup_{x \in \mathcal{H}} \langle y, x \rangle - f(x). 
\]

(2.1.4)

Let \(I_\mathcal{H}\) denote the identity map. Finally, for any \(x \in \mathcal{H}\) and scalar \(\gamma \in \mathbb{R}_{++}\), we let

\[
\text{prox}_{\gamma f}(x) := \arg\min_{y \in \mathcal{H}} f(y) + \frac{1}{2\gamma} \|y - x\|^2 
\]

\[
\text{refl}_{\gamma f} := 2\text{prox}_{\gamma f} - I_\mathcal{H},
\]

(2.1.5)

which are known as the \textit{proximal} and \textit{reflection} operators, and we define the PRS operator:

\[
T_{\text{PRS}} := \text{refl}_{\gamma f} \circ \text{refl}_{\gamma g}.
\]

(2.1.6)
2.1.3 Assumptions

We list the assumptions used throughout this chapter as follows.

**Assumption 2.1.1.** *Every function we consider is closed, proper, and convex.*

Unless otherwise stated, a function is not necessarily differentiable.

**Assumption 2.1.2 (Differentiability).** *Every differentiable function we consider is Fréchet differentiable [11, Definition 2.45].*

**Assumption 2.1.3 (Solution existence).** Functions $f, g : \mathcal{H} \to (-\infty, \infty]$ satisfy

$$\text{zer}(\partial f + \partial g) \neq \emptyset. \quad (2.1.7)$$

Note that this assumption is slightly stronger than the existence of a minimizer, because $\text{zer}(\partial f + \partial g) \neq \text{zer}(\partial(f + g))$, in general [11, Remark 16.7]. Nevertheless, this assumption is standard.

2.1.4 The Algorithms

This chapter covers several operator-splitting algorithms that are all based on the atomic evaluation of the *proximal* and *gradient* operators. By default, all algorithms start from an arbitrary $z^0 \in \mathcal{H}$. To minimize a function $f$, the *proximal point algorithm* (PPA) iteratively applies the proximal operator of $f$ as follows:

$$z^{k+1} = \text{prox}_\gamma f(z^k), \quad k = 0, 1, \ldots \quad (2.1.8)$$

where $\gamma > 0$ is a tuning parameter. Another equivalent form of the iteration, which is often used in this chapter, is

$$z^{k+1} = z^k - \gamma \tilde{\nabla} f(z^{k+1}) \quad (2.1.9)$$

where $\tilde{\nabla} f(z^{k+1}) := (1/\gamma)(z^k - z^{k+1}) \in \partial f(z^{k+1})$. Given $z^k$, the point $z^{k+1}$ is unique and so is the subgradient $\tilde{\nabla} f(z^{k+1})$ (Lemma 2.1.1). The iteration resembles the (sub)gradient
descent iteration, which uses a (sub)gradient of $f$ at $z^k$ instead of its (sub)gradient at $z^{k+1}$.

In the literature, (2.1.9) is referred to as the \textit{backward} iteration, where the (sub)gradient is drawn at the destination $z^{k+1}$. On the contrary, a \textit{forward} iteration draws the (sub)gradient at the start $z^k$, resulting in the update rule: $z^{k+1} = z^k - \gamma \nabla f(z^k)$ for an arbitrary $\nabla f(z^k) \in \partial f(z^k)$. Most of the splitting schemes in this chapter are built from forward, backward, and reflection operators.

In problem (2.1.1), let $g$ be a $C^1$ function with Lipschitz derivative. The \textit{forward-backward splitting} (FBS) algorithm is the iteration:

$$z^{k+1} = \text{prox}_{\gamma f}(z^k - \gamma \nabla g(z^k)), \quad k = 0, 1, \ldots$$  \hfill (2.1.10)

The FBS algorithm directly generalizes PPA and has the following subgradient representation:

$$z^{k+1} = z^k - \gamma \nabla f(z^{k+1}) - \gamma \nabla g(z^k)$$  \hfill (2.1.11)

where $\nabla f(z^{k+1}) := (1/\gamma)(z^k - z^{k+1} - \gamma \nabla g(z^k)) \in \partial f(z^{k+1})$, and $z^{k+1}$ and $\nabla f(z^{k+1})$ are unique (Lemma 2.1.1) given $z^k$ and $\gamma > 0$.

A direct application of the PPA algorithm (2.1.8) to minimizing $f + g$ would require computing the operator $\text{prox}_{\gamma(f+g)}$, which can be difficult to evaluate. The Douglas-Rachford splitting (DRS) algorithm eliminates this difficulty by separately evaluating the proximal operators of $f$ and $g$ as follows:

$$\begin{cases}
x_g^k = \text{prox}_{\gamma g}(z^k); \\
x_f^k = \text{prox}_{\gamma f}(2x_g^k - z^k); \quad k = 0, 1, \ldots, \\
z^{k+1} = z^k + (x_f^k - x_g^k).
\end{cases}$$

which has the equivalent operator-theoretic and subgradient form (Lemma 2.4.1):  

$$z^{k+1} = \frac{1}{2}(I_H + T_{PRS})(z^k) = z^k - \gamma(\nabla f(x_f^k) + \nabla g(x_g^k)), \quad k = 0, 1, \ldots$$
where $\tilde{\nabla} f(x^k) \in \partial f(x^k)$ and $\tilde{\nabla} g(x^k) \in \partial g(x^k)$ (see Lemma 2.4.1 for their precise definitions). In the above algorithm, we can replace the $1/2$ average of $I_H$ and $T_{PRS}$ with any other weight, so in this chapter we study the relaxed PRS algorithm:

**Algorithm 1:** Relaxed Peaceman-Rachford splitting (relaxed PRS)

```plaintext```
input : $z^0 \in H$, $\gamma > 0$, $(\lambda_j)_{j \geq 0} \subseteq (0, 1]$
for $k = 0, 1, \ldots$ do
  $z^{k+1} = (1 - \lambda_k)z^k + \lambda_k \text{refl}_{\gamma f} \circ \text{refl}_{\gamma g}(z^k)$;
```

The special cases $\lambda_k \equiv 1/2$ and $\lambda_k \equiv 1$ are called the DRS and PRS algorithms, respectively.

The relaxed PRS algorithm can be applied to problem (2.1.2). To this end we define the Lagrangian:

$$\mathcal{L}_\gamma(x, y; w) := f(x) + g(y) - \langle w, Ax + By - b \rangle + \frac{\gamma}{2} \|Ax + By - b\|^2.$$  

Section 2.8 presents Algorithm 1 applied to the Lagrange dual of (2.1.2), which reduces to the following algorithm:

**Algorithm 2:** Relaxed alternating direction method of multipliers (relaxed ADMM)

```plaintext```
input : $w^{-1} \in H$, $x^{-1} = 0$, $y^{-1} = 0$, $\lambda_{-1} = 1/2$, $\gamma > 0$, $(\lambda_j)_{j \geq 0} \subseteq (0, 1]$
for $k = -1, 0, \ldots$ do
  $y^{k+1} = \arg \min_y \mathcal{L}_\gamma(x^k, y; w^k) + \gamma(2\lambda_k - 1)(By, (Ax^k + By^k - b));$
  $w^{k+1} = w^k - \gamma(Ax^k + By^{k+1} - b) - \gamma(2\lambda_k - 1)(Ax^k + By^k - b);$
  $x^{k+1} = \arg \min_x \mathcal{L}_\gamma(x, y^{k+1}; w^{k+1});$
```

If $\lambda_k \equiv 1/2$, Algorithm 2 recovers the standard ADMM.

Each of the above algorithms is a special case of the Krasnosel’skiǐ-Mann (KM) iteration [79, 88]. An *averaged operator* is the average of a nonexpansive operator $T : H \to H$ and the identity mapping $I_H$. In other words, for all $\lambda \in (0, 1)$, the operator

$$T_\lambda := (1 - \lambda)I_H + \lambda T$$  

(2.1.12)

is called $\lambda$-averaged and every $\lambda$-averaged operator is exactly of the form $T_\lambda$ for some
nonexpansive map $T$.

Given a nonexpansive map $T$, the fixed-point iteration of the map $T_{\lambda}$ is called the KM algorithm:

**Algorithm 3: Krasnosel’skiĭ-Mann (KM)**

**input**: $z^0 \in \mathcal{H}, (\lambda_j)_{j \geq 0} \subseteq (0, 1]$

**for** $k = 0, 1, \ldots$ **do**

$z^{k+1} = T_{\lambda_k}(z^k)$;

---

### 2.1.5 Basic properties of averaged operators

This section describes the basic properties of proximal, reflection, nonexpansive, and averaged operators. We demonstrate that proximal and reflection operators are nonexpansive maps, and that averaged operators have a contractive property. These properties are included in textbooks such as [11].

**Lemma 2.1.1** (Optimality conditions of $\text{prox}$). Let $x \in \mathcal{H}$. Then $x^+ = \text{prox}_{\gamma f}(x)$ if, and only if, $(1/\gamma)(x - x^+) \in \partial f(x^+)$. It is straightforward to use Lemma 2.1.1 to deduce the firm nonexpansiveness of the proximal operator.

**Proposition 2.1.1** (Firm nonexpansiveness of $\text{prox}$). Let $x, y \in \mathcal{H}$, let $x^+ := \text{prox}_{\gamma f}(x)$, and let $y^+ := \text{prox}_{\gamma f}(y)$. Then

$$
\|x^+ - y^+\|^2 \leq \langle x^+ - y^+, x - y \rangle. \tag{2.1.13}
$$

In particular, $\text{prox}_{\gamma f}$ is nonexpansive.

The next proposition introduces the most important operator in this chapter.

**Proposition 2.1.2** (Nonexpansiveness of the PRS operator). The operator $\text{refl}_{\gamma f} : \mathcal{H} \to \mathcal{H}$ is nonexpansive. Therefore, the composition is nonexpansive:

$$
T_{\text{PRS}} := \text{refl}_{\gamma f} \circ \text{refl}_{\gamma g} \tag{2.1.14}
$$
The next proposition shows that averaged operators have a nice contraction property.

**Proposition 2.1.3** (Contraction property of averaged operator). Let $T : \mathcal{H} \to \mathcal{H}$ be a nonexpansive operator. Then for all $\lambda \in (0, 1]$ and $(x, y) \in \mathcal{H} \times \mathcal{H}$, the averaged operator $T_\lambda$ defined in (2.1.12) satisfies
\[
\|T_\lambda x - T_\lambda y\|^2 \leq \|x - y\|^2 - \frac{1 - \lambda}{\lambda}\|(I_\mathcal{H} - T_\lambda)x - (I_\mathcal{H} - T_\lambda)y\|^2. \tag{2.1.15}
\]

Note that an operator $N : \mathcal{H} \to \mathcal{H}$ satisfies the property in Equation (2.1.15) (with $N$ in place of $T_\lambda$) if, and only if, it is $\lambda$-averaged. If $\lambda = 1/2$, then $T_\lambda$ is called firmly nonexpansive. Rearranging Equation (2.1.15) shows that a nonexpansive operator $T$ is firmly nonexpansive, if, and only if, for all $x, y \in \mathcal{H}$, the inequality holds:
\[
\|Tx - Ty\|^2 \leq \langle Tx - Ty, x - y \rangle.
\]

The next corollary applies Proposition 2.1.3 to $\text{prox}_{\gamma f}$.

**Corollary 2.1.1** (Proximal operators are 1/2-averaged). The operator $\text{prox}_{\gamma f} : \mathcal{H} \to \mathcal{H}$ is 1/2-averaged and satisfies the following contraction property:
\[
\|\text{prox}_{\gamma f}(x) - \text{prox}_{\gamma f}(y)\|^2 \leq \|x - y\|^2 - \|(x - \text{prox}_{\gamma f}(x)) - (y - \text{prox}_{\gamma f}(y))\|^2. \tag{2.1.16}
\]

The following lemma relates the fixed points of $T_\lambda$ to those of $T$.

**Lemma 2.1.2.** Let $T : \mathcal{H} \to \mathcal{H}$ be nonexpansive and $\lambda > 0$. Then, $T_\lambda$ and $T$ have the same set of fixed points.

Finally, we note that the forward and forward-backward operators are averaged whenever the implicit stepsize parameter $\gamma$ is small enough. See Section 2.3.3 for more details.

### 2.2 Summable sequence convergence lemma

This section presents a lemma on the convergence rates of nonnegative summable sequences. Such sequences are constructed throughout this chapter to establish various rates.
Lemma 2.2.1 (Summable sequence convergence rates). Suppose that the nonnegative scalar sequences \((\lambda_j)_{j \geq 0}\) and \((a_j)_{j \geq 0}\) satisfy \(\sum_{i=0}^{\infty} \lambda_i a_i < \infty\). Let \(\Lambda_k := \sum_{i=0}^{k} \lambda_i\) for \(k \geq 0\).

1. **Monotonicity:** If \((a_j)_{j \geq 0}\) is monotonically nonincreasing, then

\[
a_k \leq \frac{1}{\Lambda_k} \left( \sum_{i=0}^{\infty} \lambda_i a_i \right) \quad \text{and} \quad a_k = o\left( \frac{1}{\Lambda_k - \Lambda_{\lceil k/2 \rceil}} \right). \tag{2.2.1}
\]

In particular,

(a) if \((\lambda_j)_{j \geq 0}\) is bounded away from 0, then \(a_k = o(1/(k+1))\);

(b) if \(\lambda_k = (k+1)^p\) for some \(p \geq 0\) and all \(k \geq 1\), then \(a_k = o(1/(k+1)^{p+1})\);

(c) as a special case, if \(\lambda_k = (k+1)\) for all \(k \geq 0\), then \(a_k = o(1/(k+1)^2)\).

2. **Monotonicity up to errors:** Let \((e_j)_{j \geq 0}\) be a sequence of scalars. Suppose that \(a_{k+1} \leq a_k + e_k\) for all \(k\) (where \(e_k\) represents an error) and that \(\sum_{i=0}^{\infty} \Lambda_i e_i < \infty\). Then

\[
a_k \leq \frac{1}{\Lambda_k} \left( \sum_{i=0}^{\infty} \lambda_i a_i + \sum_{i=0}^{\infty} \Lambda_i e_i \right) \quad \text{and} \quad a_k = o\left( \frac{1}{\Lambda_k - \Lambda_{\lceil k/2 \rceil}} \right). \tag{2.2.2}
\]

The rates of \(a_k\) in Parts 1a, 1b, and 1c continue to hold as long as \(\sum_{i=0}^{\infty} \Lambda_i e_i < \infty\) holds. In particular, they hold if \(e_k = O(1/(k+1)^q)\) for some \(q > 2\), \(q > p+2\), and \(q > 3\), respectively.

3. **Faster rates:** Suppose \((b_j)_{j \geq 0}\) and \((e_j)_{j \geq 0}\) are nonnegative scalar sequences, that \(\sum_{i=0}^{\infty} b_j < \infty\) and \(\sum_{i=0}^{\infty} (i+1) e_i < \infty\), and that for all \(k \geq 0\) we have \(\lambda_k a_k \leq b_k - b_{k+1} + e_k\). Then the following sum is finite:

\[
\sum_{i=0}^{\infty} (i+1) \lambda_i a_i \leq \sum_{i=0}^{\infty} b_i + \sum_{i=0}^{\infty} (i+1) e_i < \infty. \tag{2.2.3}
\]

In particular,

(a) if \((\lambda_j)_{j \geq 0}\) is bounded away from 0, then \(a_k = o(1/(k+1)^2)\);

(b) if \(\lambda_k = (k+1)^p\) for some \(p \geq 0\) and all \(k \geq 1\), then \(a_k = o(1/(k+1)^{p+2})\).
4. **No monotonicity:** For all $k \geq 0$, define the sequence of indices

$$ k_{\text{best}} := \arg \min_i \{a_i | i = 0, \ldots, k\}. $$

Then $(a_{k_{\text{best}}})_{j \geq 0}$ is monotonically nonincreasing and the above bounds continue to hold when $a_k$ is replaced with $a_{k_{\text{best}}}$.

**Proof.** Part 1. Because $a_k \leq a_i, \forall k \geq i$, and the inequality holds $\lambda_i a_i \geq 0$, we get the upper bound $\Lambda_k a_k \leq \sum_{i=0}^{k} \lambda_i a_i \leq \sum_{i=0}^{\infty} \lambda_i a_i$. This shows the left part of (2.2.1). To prove the right part of (2.2.1), observe that

$$ (\Lambda_k - \Lambda_{\lceil k/2 \rceil}) a_k = \sum_{i=\lceil k/2 \rceil}^{k} \lambda_i a_k \leq \sum_{i=\lceil k/2 \rceil}^{k} \lambda_i a_i \xrightarrow{k \to \infty} 0. $$

**Part 1a.** Let $\Lambda := \inf_{j \geq 0} \lambda_j > 0$. For every integer $k \geq 2$, we have $\lceil k/2 \rceil \leq (k + 1)/2$. Thus, $\Lambda_k - \Lambda_{\lceil k/2 \rceil} \geq \Lambda(k - \lceil k/2 \rceil) \geq \Lambda(k - 1)/2 \geq \Lambda(k + 1)/6$. Hence, $a_k = o(1/(\Lambda_k - \Lambda_{\lceil k/2 \rceil})) = o(1/(k + 1))$ follows from (2.2.1).

**Part 1b.** For every integer $k \geq 3$, we have $\lceil k/2 \rceil + 1 \leq (k + 3)/2 \leq 3(k + 1)/4$ and $\Lambda_k - \Lambda_{\lceil k/2 \rceil} = \sum_{i=\lceil k/2 \rceil + 1}^{k} \lambda_i = \sum_{i=\lceil k/2 \rceil + 1}^{k} (i + 1)^p \geq \int_{\lceil k/2 \rceil}^{k} (t + 1)^p dt = (p + 1)^{-1}((k + 1)^{p+1} - ([k/2] + 1)^{p+1}) \geq (p + 1)^{-1}(1 - (3/4)^{p+1})(k + 1)^{p+1}$. Therefore, $a_k = o(1/(\Lambda_k - \Lambda_{\lceil k/2 \rceil})) = o(1/(k + 1)^{p+1})$ follows from (2.2.1).

**Part 1c.** directly follows from Part 1b.

**Part 2.** For every integer $0 \leq i \leq k$, we have $a_k \leq a_i + \sum_{j=i}^{k-1} e_j$. Thus, $\Lambda_k a_k = \sum_{i=0}^{k} \lambda_i a_k \leq \sum_{i=0}^{k} \lambda_i a_i + \sum_{i=0}^{k} \lambda_i \sum_{j=i}^{k-1} e_j = \sum_{i=0}^{k} \lambda_i a_i + \sum_{i=0}^{k} e_i \sum_{j=0}^{k-1} \lambda_j = \sum_{i=0}^{k} \lambda_i a_i + \sum_{i=0}^{k} \lambda_i e_i \leq \sum_{i=0}^{\infty} \lambda_i a_i + \sum_{i=0}^{\infty} \lambda_i e_i$, from which the left part of (2.2.2) follows. The proof for the right part of (2.2.2) is similar to Part 1. The condition $e_k = O(1/(k + 1)^q)$ for appropriate $q$ is used to ensure that $\sum_{i=0}^{\infty} \lambda_i e_i < \infty$ for each setting of $\lambda_k$ in the previous Parts 1a, 1b, and 1c.

**Part 3.** Note that

$$ \lambda_k(k + 1)a_k \leq (k + 1)b_k - (k + 1)b_{k+1} + (k + 1)e_k $$

$$ = b_{k+1} + ((k + 1)b_k - (k + 2)b_{k+1}) + (k + 1)e_k. $$
Thus, because the upper bound on \((k + 1)\lambda_k a_k\) is the sum of a telescoping term and a summable term, we have 
\[
\sum_{i=0}^{\infty} (i + 1)\lambda_i a_i \leq \sum_{i=0}^{\infty} b_i + \sum_{i=0}^{\infty} (i + 1)e_i < \infty.
\]
Parts 3a and 3b are similar to Part 1b.

Part 4 is straightforward, so we omit its proof.

Part 1 of Lemma 2.2.1 is a generalization of [77, Theorem 3.3.1] and [57, Lemma 1.2], which state that a nonnegative, summable, monotonic sequence converges at the rate of \(o(1/(k + 1))\). This result is key for deducing the convergence rates of several quantities in this chapter.

2.3 Iterative fixed-point residual analysis

In this section we establish the convergence rate of the fixed-point residual (FPR), \(\|Tz^k - z^k\|^2\), at the \(k\)th iteration of Algorithm 3.

The convergence of Algorithm 3 is well-studied [43, 51, 82]. In particular, weak convergence of \((z^j)_{j \geq 0}\) to a fixed point of \(T\) holds under mild conditions on the sequence \((\lambda_j)_{j \geq 0}\) [43, Theorem 3.1]. Because strong convergence of Algorithm 3 may fail (in the infinite dimensional setting), the quantity \(\|z^k - z^*\|\) where \(z^*\) is a fixed point of \(T\) may be bounded above zero for all \(k \geq 0\). However, the property \(\lim_{k \to \infty} \|Tz^k - z^k\| = 0\), known as asymptotic regularity [31], always holds when a fixed point of \(T\) exists. Thus, we can always measure the convergence rate of the FPR.

We measure \(\|Tz^k - z^k\|^2\) when we could just as well measure \(\|Tz^k - z^k\|\). We choose to measure the squared norm because it naturally appears in our analysis. In addition, it is summable and monotonic, which is analyzable by Lemma 2.2.1.

In first-order optimization algorithms, the FPR typically relates to the size of objective gradient. For example, in the unit-step gradient descent algorithm, \(z^{k+1} = z^k - \nabla f(z^k)\), the FPR is given by \(\|\nabla f(z^k)\|^2\). In the proximal point algorithm, the FPR is given by
\[ \| \tilde{\nabla} f(z^{k+1}) \|_2 \text{ where } \tilde{\nabla} f(z^{k+1}) := (z^k - z^{k+1}) \in \partial f(z^{k+1}) \text{ (see Equation (2.1.9))}. \] When the objective is the sum of multiple functions, the FPR is a combination of the (sub)gradients of those functions in the objective. Using the subgradient inequality, we will derive a rate on \( f(z^k) - f(x^*) \) from a rate on the FPR where \( x^* \) is a minimizer of \( f \).

2.3.1 \( o(1/(k+1)) \) FPR of averaged operators

We now prove the main result of this section. We do not include the known weak convergence result [43, Theorem 3.1], but we deduce a convergence rate for the FPR. The new results in the following theorem are the little-\( o \) convergence rates in Equation (2.3.3) and in Part 5; the rest of the results can be found in [11, Proof of Proposition 5.14], [51, Proposition 11], and [82].

**Theorem 2.3.1** (Convergence rate of averaged operators). Let \( T : \mathcal{H} \to \mathcal{H} \) be a nonexpansive operator, let \( z^* \) be a fixed point of \( T \), let \( (\lambda_j)_{j \geq 0} \subseteq (0, 1] \) be a sequence of positive numbers, let \( \tau_k := \lambda_k(1 - \lambda_k) \), and let \( z^0 \in \mathcal{H} \). Suppose that \( (z^i)_{j \geq 0} \subseteq \mathcal{H} \) is generated by Algorithm 3: for all \( k \geq 0 \), let

\[ z^{k+1} = T\lambda_k(z^k), \tag{2.3.1} \]

where \( T\lambda \) is defined in (2.1.12). Then, the following results hold

1. \( \| z^k - z^* \|_2 \) is monotonically nonincreasing;
2. \( \| T z^k - z^k \|_2 \) is monotonically nonincreasing;
3. \( \tau_k \| T z^k - z^k \|_2 \) is summable:
\[
\sum_{i=0}^{\infty} \tau_i \| T z^i - z^i \|_2 \leq \| z^0 - z^* \|_2; \tag{2.3.2}
\]
4. if \( \tau_k > 0 \) for all \( k \geq 0 \), then the convergence estimates hold:
\[
\| T z^k - z^k \|_2 \leq \frac{\| z^0 - z^* \|_2}{\sum_{i=0}^{k} \tau_i} \quad \text{and} \quad \| T z^k - z^k \|_2 = o \left( \frac{1}{\sum_{i=|\frac{k}{2}|+1}^{k} \tau_i} \right). \tag{2.3.3}
\]

In particular, if \( (\tau_j)_{j \geq 0} \subseteq (\varepsilon, \infty) \) for some \( \varepsilon > 0 \), then \( \| T z^k - z^k \|_2 = o (1/(k+1)) \).
5. Instead of Iteration (2.3.1), for all \( k \geq 0 \), let
\[
z^{k+1} := T_{\lambda_k}(z^k) + \lambda_k e^k
\] (2.3.4)
for an error sequence \( (e^j)_{j \geq 0} \subseteq \mathcal{H} \) that satisfies \( \sum_{i=0}^{k} \lambda_i \|e_i\| < \infty \) and \( \sum_{i=0}^{\infty} (i + 1)\lambda^2_i \|e_i\|^2 < \infty \). Note that these bounds hold, for example, when for all \( k \geq 0 \)
\( \lambda_k \|e_k\| \leq \omega_k \) for a sequence \( (\omega_j)_{j \geq 0} \) that is nonnegative, summable, and monotonically nonincreasing. Then if \( (\tau_j)_{j \geq 0} \subseteq (\varepsilon, \infty) \) for some \( \varepsilon > 0 \), we continue to have \( \|Tz^k - z^k\|^2 = o(1/(k + 1)) \).

Proof. As noted before the Theorem, for Parts 1 through 4, we only need to prove the little-o convergence rate. This follows from the monotonicity of \( (\|Tz^i - z^i\|^2)_{j \geq 0} \), Equation (2.3.2), and Part 1 of Lemma 2.2.1.

Part 5: We first show that the condition involving the sequence \( (w_j)_{j \geq 0} \) is sufficient to guarantee the error bounds. We have \( \sum_{i=0}^{\infty} \lambda_i \|e_i\|^2 \leq \sum_{i=0}^{\infty} \omega_i < \infty \) and \( \sum_{i=0}^{\infty} (i + 1)\lambda^2_i \|e_i\|^2 \leq \sum_{i=0}^{\infty} (i + 1)\omega_i^2 < \infty \), where the last inequality is shown as follows. By Part 1 of Lemma 2.2.1, we have \( \omega_k = o(1/(k + 1)) \). Therefore, there exists a finite \( K \) such that \( (k + 1)\omega_k < 1 \) for \( k > K \). Therefore, \( \sum_{i=0}^{\infty} (i + 1)\omega_i^2 < \sum_{i=0}^{K} (i + 1)\omega_i^2 + \sum_{i=K+1}^{\infty} \omega_i < \infty \).

For simplicity, introduce \( p^k := Tz^k - z^k \), \( p^{k+1} := Tz^{k+1} - z^{k+1} \), and \( r^k := z^{k+1} - z^k \). Then from (2.1.12) and (2.3.4), we have \( p^k = \frac{1}{\lambda_k} (r^k - \lambda_k e^k) \). Also introduce \( q^k := Tz^{k+1} - Tz^k \). Then, \( p^{k+1} - p^k = q^k - r^k \).

We will show: (i) \( \|p^{k+1}\|^2 \leq \|p^k\|^2 + \frac{\lambda^2_k}{\tau_k} \|e^k\|^2 \) and (ii) \( \sum_{i=0}^{\infty} \tau_i \|p^i\|^2 < \infty \). Then, applying Part 2 of Lemma 2.2.1 (with \( a_k = \|p^k\|^2 \), \( e_k = \frac{\lambda^2_k}{\tau_k} \|e^k\|^2 \), and \( \lambda_k = 1 \) for which we have \( \Lambda_k = \sum_{i=0}^{k} \lambda_i \leq (k + 1) \)), we immediately obtain the rate \( \|Tz^k - z^k\|^2 = o(1/(k + 1)) \).

To prove (i), we have
\[
\|p^{k+1}\|^2 = \|p^k\|^2 + \|p^{k+1} - p^k\|^2 + 2\langle p^{k+1} - p^k, p^k \rangle
\]
\[
= \|p^k\|^2 + \|q^k - r^k\|^2 + 2\frac{1}{\lambda_k} \langle q^k - r^k, r^k - \lambda_k e^k \rangle.
\]
By the nonexpansiveness of \( T \), we have \( \|q^k\|^2 \leq \|r^k\|^2 \) and thus
\[
2\langle q^k - r^k, r^k \rangle = \|q^k\|^2 - \|r^k\|^2 - \|q^k - r^k\|^2 \leq -\|q^k - r^k\|^2.
\]
Therefore,

\[ \|p^{k+1}\|^2 \leq \|p^k\|^2 - \frac{1 - \lambda_k}{\lambda_k} \|q^k - r^k\|^2 + 2\langle q^k - r^k, e^k \rangle. \]

\[ = \|p^k\|^2 - \frac{1 - \lambda_k}{\lambda_k} \left\| q^k - r^k - \frac{\lambda_k}{1 - \lambda_k} e^k \right\|^2 + \frac{\lambda_k}{1 - \lambda_k} \|e^k\|^2 \]
\[ \leq \|p^k\|^2 + \frac{\lambda_k^2}{\tau_k} \|e^k\|^2. \]

To prove (ii): First, \(\|z^k - z^*\|\) is uniformly bounded because \(\|z^{k+1} - z^*\| \leq (1 - \lambda_k)\|z^k - z^*\| + \lambda_k\|Tz^k - z^*\| + \lambda_k\|e^k\| \leq \|z^k - z^*\| + \lambda_k\|e^k\|\) by the triangle inequality and the nonexpansiveness of \(T\). From [11, Corollary 2.14], we have

\[ \|z^{k+1} - z^*\|^2 = \|(1 - \lambda_k)(z^k - z^*) + \lambda_k(Tz^k - z^* + e^k)\|^2 \]
\[ = (1 - \lambda_k)\|z^k - z^*\|^2 + \lambda_k\|Tz^k - z^* + e^k\|^2 - \lambda_k(1 - \lambda_k)\|p^k + e^k\|^2 \]
\[ = (1 - \lambda_k)\|z^k - z^*\|^2 + \lambda_k \left( \|Tz^k - z^*\|^2 + 2\lambda_k \langle Tz^k - z^*, e^k \rangle + \lambda_k\|e^k\|^2 \right) \]
\[ - \lambda_k(1 - \lambda_k) \left( \|p^k\|^2 + 2\langle p^k, e^k \rangle + \|e^k\|^2 \right) \]
\[ \leq \|z^k - z^*\|^2 - \tau_k\|p^k\|^2 + \lambda_k^2\|e^k\|^2 + 2\lambda_k\|Tz^k - z^*\|\|e^k\| + 2\tau_k\|p^k\|\|e^k\| \]
\[ = \|z^k - z^*\|^2 - \tau_k\|p^k\|^2 + \xi_k. \]

Because we have shown (a) \(\|Tz^k - z^*\|\) and \(\|p^k\|\) are bounded, (b) \(\sum_{k=0}^{\infty} \lambda_k\|e^k\| \leq \sum_{k=0}^{\infty} \lambda_k\|e^k\| < \infty\), and (c) \(\sum_{k=0}^{\infty} \lambda_k^2\|e^k\|^2 \leq \infty\), we have \(\sum_{i=0}^{\infty} \xi_k < \infty\) and thus \(\sum_{i=0}^{\infty} \tau_k\|Tz_i - z^*\|^2 \leq \|z^0 - z^*\|^2 + \sum_{i=0}^{\infty} \xi_k < \infty. \)

\[ \square \]

2.3.1.1 Notes on Theorem 2.3.1

The FPR, \(\|Tz^k - z^k\|^2\), is a normalized version of the successive iterate differences \(z^{k+1} - z^k = \lambda_k(Tz^k - z^k)\). Thus, the convergence rates of \(\|Tz^k - z^k\|^2\) naturally induce convergence rates of \(\|z^{k+1} - z^k\|^2\).

Note that \(o(1/(k+1))\) is the optimal convergence rate for the class of nonexpansive operators [28, Remarque 4]. In the special case that \(T = \text{prox}_{\gamma f}\) for some closed, proper, and convex function \(f\), the rate of \(\|Tz^k - z^k\|^2\) improves to \(O(1/(k+1)^2)\) [28, Théorème
See section 2.6 for more optimality results. Also, the little-o convergence rate of the fixed-point residual associated to the resolvent of a maximal monotone linear operator was shown in [28, Proposition 4]. Finally, we mention the parallel work [53], which proves a similar little-o convergence rate for the fixed-point residual of relaxed PPA.

In general, it is possible that the nonexpansive operator, $T : \mathcal{H} \to \mathcal{H}$, is already averaged, i.e. there exists a nonexpansive operator $N : \mathcal{H} \to \mathcal{H}$ and a positive constant $\alpha \in (0, 1]$ such that $T = (1 - \alpha)I_H + \alpha N$. In this case, Lemma 2.1.2 shows that $T$ and $N$ share the same fixed point set. Thus, we can apply Theorem 2.3.1 to $N = (1 - (1/\alpha))I_H + (1/\alpha)T$. Furthermore, $N_{\lambda} = (1 - \lambda/\alpha)I_H + (\lambda/\alpha)T$. Thus, when we translate this back to an iteration on $T$, it enlarges the region of relaxation parameters to $\lambda_k \in (0, 1/\alpha)$ and modifies $\tau_k$ accordingly to $\tau_k = \lambda_k(1 - \alpha\lambda_k)/\alpha$, and the same convergence results continue to hold.

To the best of our knowledge, the little-o rates produced in Theorem 2.3.1 have never been established for the KM iteration. See [51, 82] for similar big-O results. Note that our rate in Part 5 is strictly better than the one shown in [82], and it is given under a much weaker condition on the error. Indeed, [82] achieves an $O(1/(k + 1))$ convergence rate whenever $\sum_{i=0}^{\infty}(i + 1)\|e^k\| < \infty$, which implies that $\min_{i=0,...,k}\{\|e^k\|\} = o(1/(k + 1)^2)$ by Lemma 2.2.1. In contrast, any error sequence of the form $\|e^k\| = O(1/(k + 1)^{\alpha})$ with $\alpha > 1$ will satisfy Part 5 of our Theorem 2.3.1. Finally, note that in the Banach space case, we cannot improve the big-O rates to little-o [51, Section 2.4].

### 2.3.2 $o(1/(k + 1))$ FPR of relaxed PRS

In this section, we apply Theorem 2.3.1 to the $T_{\text{PRS}}$ operator defined in Proposition 2.1.2. For the special case of DRS ($(1/2)$-averaged PRS), it is straightforward to establish the rate of the FPR

$$\|(T_{\text{PRS}})_{1/2}z^k - z^k\|^2 = O\left(\frac{1}{k + 1}\right)$$
from two existing results: (i) the DRS iteration is a proximal iteration applied to a certain monotone operator [60, Section 4]; (ii) the convergence rate of the FPR for proximal iterations is $O(1/(k + 1))$ [28, Proposition 8] whenever a fixed point exists. Our results below are established for general averaged PRS operators and the rate is improved to $o(1/(k + 1))$.

The following corollary is an immediate consequence of Theorem 2.3.1.

**Corollary 2.3.1** (Convergence rate of relaxed PRS). Let $z^*$ be a fixed point of $T_{PRS}$, let $(\lambda_j)_{j \geq 0} \subseteq (0, 1]$ be a sequence of positive numbers, let $\tau_k := \lambda_k(1 - \lambda_k)$ for all $k \geq 0$, and let $z^0 \in \mathcal{H}$. Suppose that $(z^i)_{j \geq 0} \subseteq \mathcal{H}$ is generated by Algorithm 1. Then the sequence $\|z^k - z^*\|^2$ is monotonically nonincreasing and the following inequality holds:

$$\sum_{i=0}^{\infty} \tau_i \|T_{PRS}z^i - z^i\|^2 \leq \|z^0 - z^*\|^2.$$  

(2.3.5)

Furthermore, if $\tau := \inf_{j \geq 0} \tau_j > 0$, then the following convergence rates hold:

$$\|T_{PRS}z^k - z^k\|^2 \leq \frac{\|z^0 - z^*\|^2}{\tau(k + 1)} \quad \text{and} \quad \|T_{PRS}z^k - z^k\|^2 = o\left(\frac{1}{\tau(k + 1)}\right).$$  

(2.3.6)

**2.3.3 $o(1/(k + 1)^2)$ FPR of FBS and PPA**

In this section, we assume that $\nabla g$ is $(1/\beta)$-Lipschitz, and we analyze the convergence rate of FBS algorithm given in Equations (2.1.10) and (2.1.11). If $g = 0$, FBS reduces to PPA and $\beta = \infty$. If $f = 0$, FBS reduces to gradient descent. The FBS algorithm can be written in the following operator form:

$$T_{FBS} := \text{prox}_{\gamma f} \circ (I - \gamma \nabla g).$$

Because $\text{prox}_{\gamma f}$ is $(1/2)$-averaged and $I - \gamma \nabla g$ is $\gamma/(2\beta)$-averaged [98, Theorem 3(b)], it follows that $T_{FBS}$ is $\alpha_{FBS}$-averaged for

$$\alpha_{FBS} := \frac{2\beta}{4\beta - \gamma} \in (1/2, 1)$$

whenever $\gamma < 2\beta$ [11, Proposition 4.32]. Thus, we have $T_{FBS} = (1 - \alpha_{FBS})I + \alpha_{FBS}T$ for a certain nonexpansive operator $T$, and $T_{FBS}(z^k) - z^k = \alpha_{FBS}(Tz^k - z^k)$. In particular,
for all $\gamma < 2\beta$ the following sum is finite:

$$\sum_{i=0}^{\infty} \|T_{FBS}(z^k) - z^k\|^2 \leq \frac{\alpha_{FBS}\|z^0 - z^*\|^2}{(1 - \alpha_{FBS})}.$$  

To analyze the FBS algorithm we need to derive a joint subgradient inequality for $f + g$. First, we recall the following sufficient descent property for Lipschitz differentiable functions.

**Theorem 2.3.2** (Descent theorem [11, Theorem 18.15(iii)]). If $g$ is differentiable and $\nabla g$ is $(1/\beta)$-Lipschitz, then for all $x, y \in \mathcal{H}$ we have the upper bound

$$g(x) \leq g(y) + \langle x - y, \nabla g(y) \rangle + \frac{1}{2\beta}\|x - y\|^2. \quad (2.3.7)$$

**Corollary 2.3.2** (Joint descent theorem). If $g$ is differentiable and $\nabla g$ is $(1/\beta)$-Lipschitz, then for all points $x, y \in \text{dom}(f)$ and $z \in \mathcal{H}$, and subgradients $\tilde{\nabla} f(x) \in \partial f(x)$, we have

$$f(x) + g(x) \leq f(y) + g(y) + \langle x - y, \nabla g(z) + \tilde{\nabla} f(x) \rangle + \frac{1}{2\beta}\|z - x\|^2. \quad (2.3.8)$$

**Proof.** Inequality (2.3.8) follows from adding the upper bound

$$g(x) - g(y) \leq g(z) - g(y) + \langle x - z, \nabla g(z) \rangle + \frac{1}{2\beta}\|z - x\|^2$$

$$\leq \langle x - y, \nabla g(z) \rangle + \frac{1}{2\beta}\|z - x\|^2$$

with the subgradient inequality: $f(x) \leq f(y) + \langle x - y, \tilde{\nabla} f(x) \rangle$.

We now improve the $O(1/(k + 1)^2)$ FPR rate for PPA in [28, Théorème 9] by showing that the FPR rate of FBS is actually $o(1/(k + 1)^2)$.

**Theorem 2.3.3** (Objective and FPR convergence of FBS). Let $z^0 \in \text{dom}(f) \cap \text{dom}(g)$ and let $x^*$ be a minimizer of $f + g$. Suppose that $(z^j)_{j \geq 0}$ is generated by FBS (iteration
(2.1.10) where \( \nabla g \) is \((1/\beta)\)-Lipschitz and \( \gamma < 2\beta \). Then for all \( k \geq 0 \),
\[
f(z^{k+1}) + g(z^{k+1}) - f(x^*) - g(x^*)
\leq \frac{\|z^0 - x^*\|^2}{k+1} \times \left\{
\frac{1}{2\gamma} \quad \text{if } \gamma \leq \beta;
\left( \frac{1}{2\gamma} + \left( \frac{1}{2\beta} - \frac{1}{2\gamma} \right) \frac{\alpha_{\text{FBS}}}{(1-\alpha_{\text{FBS}})} \right) \quad \text{otherwise},
\right.\]
and
\[
f(z^{k+1}) + g(z^{k+1}) - f(x^*) - g(x^*) = o(1/(k+1)).
\]
In addition, for all \( k \geq 0 \), we have \( \|T_{\text{FBS}}z^{k+1} - z^{k+1}\|^2 = o(1/(k+1)^2) \) and
\[
\|T_{\text{FBS}}z^{k+1} - z^{k+1}\|^2 \leq \frac{\|z^0 - x^*\|^2}{(1/\gamma - 1/2\beta)(k+1)^2} \times \left\{
\frac{1}{2\gamma} \quad \text{if } \gamma \leq \beta;
\left( \frac{1}{2\gamma} + \left( \frac{1}{2\beta} - \frac{1}{2\gamma} \right) \frac{\alpha_{\text{FBS}}}{(1-\alpha_{\text{FBS}})} \right) \quad \text{otherwise}.
\right.
\]

**Proof.** Recall that \( z^k - z^{k+1} = \gamma \tilde{\nabla} f(z^{k+1}) + \gamma \nabla g(z^k) \in \gamma \partial f(z^{k+1}) + \gamma \nabla g(z^k) \) for all \( k \geq 0 \). Thus, the joint descent theorem shows that for all \( x \in \text{dom}(f) \), we have
\[
f(z^{k+1}) + g(z^{k+1}) - f(x) - g(x)
\leq \frac{1}{\gamma} \langle z^{k+1} - x, z^k - z^{k+1} \rangle + \frac{1}{2\beta} \|z^k - z^{k+1}\|^2
= \frac{1}{2\gamma} \left( \|z^k - x\|^2 - \|z^{k+1} - x\|^2 \right) + \left( \frac{1}{2\beta} - \frac{1}{2\gamma} \right) \|z^{k+1} - z^k\|^2. \tag{2.3.9}
\]
Let \( h := f + g \). If we set \( x = x^* \) in Equation (2.3.9), we see that \( (h(z^{j+1}) - h(x^*))_{j \geq 0} \) is positive, summable, and
\[
\sum_{i=0}^{\infty} (h(z^{i+1}) - h(x^*)) \leq \left\{
\frac{1}{2\gamma} \|z^0 - x^*\|^2 \quad \text{if } \gamma \leq \beta;
\left( \frac{1}{2\gamma} + \left( \frac{1}{2\beta} - \frac{1}{2\gamma} \right) \frac{\alpha_{\text{FBS}}}{(1-\alpha_{\text{FBS}})} \right) \|z^0 - x^*\|^2 \quad \text{otherwise},
\right.\] \tag{2.3.10}

In addition, if we set \( x = z^k \) in Equation (2.3.9), then we see that \( (h(z^{j+1}) - h(x^*))_{j \geq 0} \) is decreasing:
\[
\left( \frac{1}{\gamma} - \frac{1}{2\beta} \right) \|z^{k+1} - z^k\|^2 \leq h(z^k) - h(z^{k+1}) = (h(z^k) - h(x^*)) - (h(z^{k+1}) - h(x^*)).
\]

23
Therefore, the rates for $f(z^{k+1}) + g(z^{k+1}) - f(x^*) - g(x^*)$ follow by Lemma 2.2.1 Part 1a, with $a_k = h(z^{k+1}) - h(x^*)$ and $\lambda_k \equiv 1$.

Now we prove the rates for $\|T_{\text{FBS}} z^{k+1} - z^{k+1}\|^2$. We apply Part 3 of Lemma 2.2.1 with $a_k = (1/\gamma - 1/(2\beta)) \|z^{k+2} - z^{k+1}\|^2$, $\lambda_k \equiv 1$, $e_k = 0$, and $b_k = h(z^{k+1}) - h(x^*)$ for all $k \geq 0$, to show that $\sum_{i=0}^{\infty} (i + 1)a_i$ is less than the sum in Equation (2.3.10). Part 2 of Theorem 2.3.1 shows that $(a_j)_{j \geq 0}$ is monotonically nonincreasing. Therefore, the convergence rate of $(a_j)_{j \geq 0}$ follows from Part 1b of Lemma 2.2.1.

When $f = 0$, the objective error upper bound in Theorem 2.3.3 is strictly better than the bound provided in [94, Corollary 2.1.2]. In FBS, the objective error rate is the same as the one derived in [14, Theorem 3.1], when $\gamma \in (0, \beta]$, and is the same as the one given in [27] in the case that $\gamma \in (0, 2\beta)$. The little-$o$ FPR rate is new in all cases except for the special case of PPA ($g \equiv 0$) under the condition that the sequence $(z^j)_{j \geq 0}$ strongly converges to a minimizer [69].

### 2.3.4 $o(1/(k + 1)^2)$ FPR of one dimensional DRS

Whenever the operator $(T_{\text{PRS}})_{1/2}$ is applied in $\mathbb{R}$, the convergence rate of the FPR improves to $o(1/(k + 1)^2)$.

**Theorem 2.3.4.** Suppose that $\mathcal{H} = \mathbb{R}$, and suppose that $(z^j)_{j \geq 0}$ is generated by the DRS algorithm, i.e. Algorithm 1 with $\lambda_k \equiv 1/2$. Then for all $k \geq 0$,

$$|(T_{\text{PRS}})_{1/2} z^{k+1} - z^{k+1}|^2 = \frac{|z^0 - z^*|^2}{2(k+1)^2} \quad \text{and} \quad |(T_{\text{PRS}})_{1/2} z^{k+1} - z^{k+1}|^2 = o\left(\frac{1}{(k+1)^2}\right).$$

**Proof.** Note that $(T_{\text{PRS}})_{1/2}$ is $(1/2)$-averaged, and, hence, it is the resolvent of some maximal monotone operator on $\mathbb{R}$ [11, Corollary 23.8]. Furthermore, every maximal monotone operator on $\mathbb{R}$ is the subdifferential operator of a closed, proper, and convex function [11, Corollary 22.19]. Therefore, DRS is equivalent to the proximal point algorithm applied
to a certain convex function on $\mathbf{R}$. Thus, the result follows by Theorem 2.3.2 applied to this function.

\[\square\]

2.3.5 $O(1/\Lambda_k^2)$ ergodic FPR of Fejér monotone sequences

The following definition has proved to be quite useful in the analysis of optimization algorithms [42].

**Definition 2.3.1** (Fejér monotone sequences). A sequence $(z_j)_{j \geq 0} \subseteq \mathcal{H}$ is Fejér monotone with respect to a nonempty set $C \subseteq \mathcal{H}$ if for all $z \in C$, we have $\|z^{k+1} - z\|^2 \leq \|z^k - z\|^2$.

The following fact is trivial, but allows us to deduce ergodic convergence rates of many algorithms.

**Theorem 2.3.5.** Let $(z_j)_{j \geq 0}$ be a Fejér monotone sequence with respect to a nonempty set $C \subseteq \mathcal{H}$. Suppose that $z^{k+1} - z^k = \lambda_k (x^k - y^k)$ for a sequence $((x^j, y^j))_{j \geq 0} \subseteq \mathcal{H}^2$, and a sequence of positive real numbers $(\lambda_j)_{j \geq 0}$. For all $k \geq 0$, let $x^k := (1/\Lambda_k) \sum_{i=0}^k \lambda_i x^i$, let $y^k := (1/\Lambda_k) \sum_{i=0}^k \lambda_i y^i$, and let $z^k := (1/\Lambda_k) \sum_{i=0}^k \lambda_i z^i$. Then we get the following bound for all $z \in C$:

$$\|x^k - y^k\|^2 \leq \frac{4\|z^0 - z\|^2}{\Lambda_k^2}.$$  

**Proof.** It follows directly from the inequality: $\lambda_k \|x^k - y^k\| = \left\| \sum_{i=0}^k (z^{k+1} - z^k) \right\| = \|z^{k+1} - z^0\| \leq 2\|z^0 - z\|$.

In view of Part 1 of Theorem 2.3.1, we see that any sequence $(z_j)_{j \geq 0}$ generated by Algorithm 3 is Fejér monotone with respect to the set of fixed-points of $T$. Therefore, Theorem 2.3.5 directly applies to the KM iteration in Equation (2.3.1) with the choice $x^k = Tz^k$ and $y^k = z^k$ for all $k \geq 0$.

The interested reader can proceed to Section 2.6 for several examples that show the optimality of the rates predicted in this section.
2.4 Subgradients and fundamental inequalities

We now shift the focus from operator-theoretic analysis to function minimization. This section establishes fundamental inequalities that connect the $FPR$ in Section 2.3 to the objective error of the relaxed PRS algorithm.

In first-order optimization algorithms, we only have access to (sub)gradients and function values. Consequently, the FPR at each iteration is usually some linear combination of (sub)gradients. In simple first-order algorithms, for example the (sub)gradient method, a (sub)gradient is drawn from a single point at each iteration. In splitting algorithms for problems with multiple convex functions, each function draws at subgradient at a different point. There is no natural point at which we can evaluate the entire objective function; this complicates the analysis of the relaxed PRS algorithm.

In the relaxed PRS algorithm, there are two objective functions $f$ and $g$, and the two operators $\text{refl}_{\gamma f}$ and $\text{refl}_{\gamma g}$ are calculated one after another at different points, neither of which equals $z^k$ or $z^{k+1}$. Consequently, the expression $z^k - z^{k+1}$ is more complicated, and the analysis for standard (sub)gradient iteration does not carry through.

We let $x_f$ and $x_g$ be the points where subgradients of $f$ and $g$ are drawn, respectively, and introduce a triangle diagram in Figure 2.1 for deducing the algebraic relations among points $z$, $x_f$ and $x_g$. These relations will be used frequently in our analysis. Propositions 2.4.1 and 2.4.2 use this diagram to bound the objective error in terms of the FPR. In these bounds, the objective errors of $f$ and $g$ are measured at two points $x_f$ and $x_g$ such that $x_f \neq x_g$. Later we will assume that one of the objectives is Lipschitz continuous and evaluate both functions at the same point (See Corollaries 2.5.1 and 2.5.2).

We conclude this introduction by combining the subgradient notation in Equation (2.1.3) and Lemma 2.1.1 to arrive at the expressions

$$\text{prox}_{\gamma f}(x) = x - \gamma \tilde{\nabla} f(\text{prox}_{\gamma f}(x)) \quad \text{and} \quad \text{refl}_{\gamma f}(x) = x - 2\gamma \tilde{\nabla} f(\text{prox}_{\gamma f}(x)).$$

(2.4.1)
With this notation, we can decompose the FPR at each iteration of the relaxed PRS algorithm in terms of subgradients drawn at certain points.

### 2.4.1 A subgradient representation of relaxed PRS

In this section we write the relaxed PRS algorithm in terms of subgradients. Lemma 2.4.1, Table 2.1, and Figure 2.1 summarize a single iteration of relaxed PRS.

![Figure 2.1: A single relaxed PRS iteration, from $z$ to $(T_{PRS})_\lambda(z)$.](image)

The way to read Figure 2.1 is as follows: Given input $z$, relaxed PRS takes a *backward–forward* step with respect to $g$, then takes a *backward–forward* step with respect to $f$, resulting in the point $T_{PRS}(z)$. (Refer to the discussion below (2.1.9) for the concepts of “backward” and “forward.”) Finally, it averages the input and output: $(T_{PRS})_\lambda(z) = (1 - \lambda)z + \lambda(T_{PRS}(z))$.

Lemma 2.4.1 summarizes and proves the identities depicted in Figure 2.1.

**Lemma 2.4.1.** Let $z \in \mathcal{H}$. Define auxiliary points $x_g := \text{prox}_{\gamma g}(z)$ and $x_f := \text{prox}_{\gamma f}(\text{refl}_{\gamma g}(z))$. Then the identities hold:

$$x_g = z - \gamma \nabla g(x_g) \quad \text{and} \quad x_f = x_g - \gamma \nabla g(x_g) - \gamma \nabla f(x_f).$$

(2.4.2)
\[
\begin{align*}
\text{Point} & \quad \text{Operator identity} & \quad \text{Subgradient identity} \\
\quad x^g & = \text{prox}_{\gamma g}(z^s) & = z^s - \gamma \tilde{\nabla} g(x^g) \\
\quad x^f & = \text{prox}_{\gamma f}(\text{refl}_{\gamma g}(z^s)) & = x^g - \gamma (\tilde{\nabla} g(x^g) + \tilde{\nabla} f(x^f)) \\
(T_{\text{PRS}}\lambda)(z^s) & = (1 - \lambda)z^s + \lambda T_{\text{PRS}}(z^s) & = z^s - 2\gamma \lambda (\tilde{\nabla} g(x^g) + \tilde{\nabla} f(x^f))
\end{align*}
\]

| Table 2.1: Overview of the main identities used throughout the chapter. The letter \(s\) denotes a superscript (e.g. \(s = k\) or \(s = \ast\)). The vector \(z^s \in H\) is an arbitrary input point. See Lemma 2.4.1 for a proof. |

where \(\tilde{\nabla} g(x_g) := (1/\gamma)(z - x_g) \in \partial g(x_g)\) and \(\tilde{\nabla} f(x_f) := (1/\gamma)(2x_g - z - x_f) \in \partial f(x_f)\).

In addition, each relaxed PRS step has the following representation:

\[
(T_{\text{PRS}}\lambda)(z) - z = 2\lambda(x_f - x_g) = -2\gamma \lambda (\tilde{\nabla} g(x_g) + \tilde{\nabla} f(x_f)).
\] (2.4.3)

**Proof.** Figure 2.1 provides an illustration of the identities. Equation (2.4.2) follows from \(\text{refl}_{\gamma g}(z) = 2x_g - z = x_g - \gamma \tilde{\nabla} g(x_g)\) and Equation (2.4.1). Now, we can compute \(T_{\text{PRS}}(z) - z\):

\[
T_{\text{PRS}}(z) - z \overset{(2.1.14)}{=} \text{refl}_{\gamma f}(\text{refl}_{\gamma g}(z)) - z = 2x_f - \text{refl}_{\gamma g}(z) - z = 2x_f - (2x_g - z) - z = 2(x_f - x_g).
\]

The subgradient identity in (2.4.3) follows from (2.4.2). Finally, Equation (2.4.3) follows from \((T_{\text{PRS}}\lambda)(z) - z = (1 - \lambda)z + \lambda T_{\text{PRS}}(z) - z = \lambda(T_{\text{PRS}}(z) - z)\). \qed

**2.4.2 Optimality conditions of relaxed PRS**

The following lemma characterizes the zeros of \(\partial f + \partial g\) in terms of the fixed points of the PRS operator. The intuition is the following: If \(z^*\) is a fixed point of \(T_{\text{PRS}}\), then the base of the triangle in Figure 2.1 has length zero. Thus, \(x^* := x^g = x^f\), and if we
travel around the perimeter of the triangle, we will start and begin at \( z^* \). This shows that
\[-2\gamma \tilde{\nabla} g(x^*) = 2\gamma \tilde{\nabla} f(x^*), \]
i.e. \( x^* \in \text{zer}(\partial f + \partial g) \).

**Lemma 2.4.2** (Optimality conditions of \( T_{PRS} \)). *The following identity holds:*

\[
\text{zer}(\partial f + \partial g) = \{ \text{prox}_{\gamma g}(z) \mid z \in \mathcal{H}, T_{PRS}z = z \}. \tag{2.4.4}
\]

That is, if \( z^* \) is a fixed point of \( T_{PRS} \), then \( x^* = x_g^* = x_f^* \) is a solution to Problem 2.1.1 and

\[
z^* - x^* = \gamma \tilde{\nabla} g(x^*) \in \gamma \partial g(x^*). \tag{2.4.5}
\]

**Proof.** See [11, Proposition 25.1] for the proof of Equation (2.4.4). Equation (2.4.5) follows because \( x^* = \text{prox}_{\gamma g}(z^*) \) if, and only if, \( z^* - x^* \in \gamma \partial g(x^*) \). \( \square \)

**2.4.3 Fundamental inequalities**

We now deduce inequalities on the objective function \( f + g \). In particular, we compute upper and lower bounds of the quantities \( f(x_f^k) + g(x_g^k) - g(x^*) - f(x^*) \). Note that \( x_f^k \) and \( x_g^k \) are not necessarily equal, so this quantity can be negative.

The most important properties of the inequalities we establish below are:

1. The upper fundamental inequality has a telescoping structure in \( z^k \) and \( z^{k+1} \).
2. They can be bounded in terms of \( \| z^{k+1} - z^k \|^2 \).

Properties 1 and 2 will be used to deduce ergodic and nonergodic rates, respectively.

**Proposition 2.4.1** (Upper fundamental inequality). *Let \( z \in \mathcal{H} \), let \( z^+ := (T_{PRS})_\lambda(z) \), and let \( x_f \) and \( x_g \) be defined as in Lemma 2.4.1. Then for all \( x \in \text{dom}(f) \cap \text{dom}(g) \)

\[
4\gamma \lambda(f(x_f) + g(x_g) - f(x) - g(x)) \leq \| z - x \|^2 - \| z^+ - x \|^2 + \left( 1 - \frac{1}{\lambda} \right) \| z^+ - z \|^2. \tag{2.4.6}
\]
Proof. We use the subgradient inequality and (2.4.3) multiple times in the following derivation:

\[ 4\gamma \lambda (f(x_f) + g(x_g) - f(x) - g(x)) \]

\[ \leq 4\lambda \gamma \left( \langle x_f - x, \tilde{\nabla} f(x_f) \rangle + \langle x_g - x, \tilde{\nabla} g(x_g) \rangle \right) \]

\[ = 4\lambda \gamma \left( \langle x_f - x_g, \tilde{\nabla} f(x_f) \rangle + \langle x_g - x, \tilde{\nabla} f(x_f) + \tilde{\nabla} g(x_g) \rangle \right) \]

\[ = 2 \left( \langle x_f - x_g, \tilde{\nabla} f(x_f) \rangle + \langle x - x_g, z^* - z \rangle \right) \]

\[ = 2 \langle z^* - z, x + (z - x_g + \gamma \tilde{\nabla} f(x_f)) - z \rangle \]

\[ = 2 \langle z^* - z, x + \gamma (\tilde{\nabla} g(x_g) + \tilde{\nabla} f(x_f)) - z \rangle \]

\[ = 2 \langle z^* - z, x - \frac{1}{2\lambda} (z^* - z) - z \rangle \]

\[ = \|z - x\|^2 - \|z^* - x\|^2 + \left(1 - \frac{1}{\lambda}\right) \|z^* - z\|^2. \]

Proposition 2.4.2 (Lower fundamental inequality). Let \( z^* \) be a fixed point of \( T_{\text{PRS}} \) and let \( x^* := \text{prox}_{\gamma g}(z^*) \). Then for all \( x_f \in \text{dom}(f) \) and \( x_g \in \text{dom}(g) \), the lower bound holds:

\[ f(x_f) + g(x_g) - f(x^*) - g(x^*) \geq \frac{1}{\gamma} \langle x_g - x_f, z^* - x^* \rangle. \quad (2.4.7) \]

Proof. This proof essentially follows from the subgradient inequality. Indeed, let \( \tilde{\nabla} g(x^*) = (z^* - x^*)/\gamma \in \partial g(x^*) \) and let \( \tilde{\nabla} f(x^*) = -\tilde{\nabla} g(x^*) \in \partial f(x^*) \). Then the result follows by adding the following equations:

\[ f(x_f) - f(x^*) \geq \langle x_f - x^*, \tilde{\nabla} f(x^*) \rangle, \]

\[ g(x_g) - g(x^*) \geq \langle x_g - x_f, \tilde{\nabla} g(x^*) \rangle + \langle x_f - x^*, \tilde{\nabla} g(x^*) \rangle. \]

\[ \square \]

2.5 Objective convergence rates

In this section we will prove ergodic and nonergodic convergence rates of relaxed PRS when \( f \) and \( g \) are closed, proper, and convex functions that are possibly nonsmooth.
To ease notational memory, we note that the reader may assume that \( \lambda_k = (1/2) \) for all \( k \geq 0 \). This simplification implies that \( \Lambda_k = (1/2)(k+1) \), and \( \tau_k = \lambda_k(1-\lambda_k) = (1/4) \) for all \( k \geq 0 \).

Throughout this section the point \( z^* \) denotes an arbitrary fixed point of \( T_{PRS} \), and we define a minimizer of \( f + g \) by the formula (Lemma 2.4.2):

\[
x^* = \text{prox}_{\gamma g}(z^*).
\]

The constant \((1/\gamma)\|z^* - x^*\|\) appears in the bounds of this section. This term is independent of \( \gamma \): For any fixed point \( z^* \) of \( T_{PRS} \), the point \( x^* = \text{prox}_{\gamma g}(z^*) \) is a minimizer and \( z^* - \text{prox}_{\gamma g}(z^*) = \gamma \nabla g(x^*) \in \gamma \partial g(x^*) \). Conversely, if \( x^* \in \text{zer}(\partial f + \partial g) \) and \( \nabla g(x^*) \in (-\partial f(x^*)) \cap \partial g(x^*) \), then \( z^* = x^* + \gamma \nabla g(x^*) \) is a fixed point. Note that in all of our bounds, we can always replace \((1/\gamma)\|z^* - x^*\| = \|\nabla g(x^*)\|\) by the infimum \( \inf_{z^* \in \text{Fix}(T_{PRS})} (1/\gamma)\|z^* - x^*\| \) (although the infimum might not be attained).

2.5.1 Ergodic convergence rates

In this section, we analyze the ergodic convergence of relaxed PRS. The proof follows the telescoping property of the upper and lower fundamental inequalities and an application of Jensen’s inequality.

**Theorem 2.5.1** (Ergodic convergence of relaxed PRS). For all \( k \geq 0 \), let \( \lambda_k \in (0, 1] \). Then we have the following convergence rate

\[
-2\|z^0 - z^*\|z^* - x^*\|_\gamma \Lambda_k \leq f(x^k_f) + g(x^k_g) - f(x^*) - g(x^*) \leq \frac{1}{4\gamma \Lambda_k}\|z^0 - x^*\|^2.
\]

In addition, the following feasibility bound holds:

\[
\|x^k_g - x^k_f\| \leq \frac{2\|z^0 - z^*\|_\Lambda_k}{\Lambda_k}.
\]

**Proof.** Equation (2.5.1) follows directly from Theorem 2.3.5 because \((z^j)_{j \geq 0}\) is Fejér monotone with respect to \( \text{Fix}(T) \) and for all \( k \geq 0 \), we have \( z^{k+1} - z^k = \lambda_k(x^k_f - x^k_g) \).
Recall the upper fundamental inequality from Proposition 2.4.1:

$$4\gamma k(f(x^k) + g(x^k) - f(x^*) - g(x^*))$$

$$\leq \|z^k - x^*\|^2 - \|z^{k+1} - x^*\|^2 + \left(1 - \frac{1}{\lambda_k}\right)\|z^{k+1} - z^k\|^2.$$  \hspace{1cm} (2.5.2)  

Because $\lambda_k \leq 1$, it follows that $(1 - (1/\lambda_k)) \leq 0$. Thus, we sum Equation (2.5.2) from $i = 0$ to $k$, divide by $\Lambda_k$, and apply Jensen’s inequality to get

$$\frac{1}{4\gamma \Lambda_k}(\|z^0 - x^*\|^2 - \|z^{k+1} - x^*\|^2) \geq \frac{1}{\Lambda_k} \sum_{i=0}^{k} \lambda_i (f(x^i_f) + g(x^i_g) - f(x^*) - g(x^*))$$

$$\geq f(\bar{x}_f^k) + g(\bar{x}_g^k) - f(x^*) - g(x^*).$$

The lower bound is a consequence of the fundamental lower inequality and Equation (2.5.1)

$$f(\bar{x}_f^k) + g(\bar{x}_g^k) - f(x^*) - g(x^*) \geq \frac{1}{\gamma}(\bar{x}_g^k - \bar{x}_f^k, z^* - x^*) \geq -\frac{2\|z^0 - z^*\|\|z^* - x^*\|}{\gamma \Lambda_k}.$$  \hspace{1cm} (2.5.4)

In general, $x^k_f \not\in \text{dom}(g)$ and $x^k_g \not\in \text{dom}(f)$, so we cannot evaluate $g$ at $x^k_f$ or $f$ at $x^k_g$. However, the conclusion of Theorem 2.5.1 can be improved if $f$ or $g$ is Lipschitz continuous. The following proposition gives a sufficient condition for Lipschitz continuity on a ball:

**Proposition 2.5.1** (Lipschitz continuity on a ball). Suppose that $f : \mathcal{H} \to (-\infty, \infty]$ is proper and convex. Let $\rho > 0$ and let $x_0 \in \mathcal{H}$. If $\delta = \sup_{x,y \in B(x_0, 2\rho)} |f(x) - f(y)| < \infty$, then $f$ is $(\delta/\rho)$-Lipschitz on $B(x_0, \rho)$.

**Proof.** See [11, Proposition 8.28].

To use this fact, we need to show that the sequences $(x^j_f)_{j \geq 0}$ and $(x^j_g)_{j \geq 0}$ are bounded. Recall that $x^s_g = \text{prox}_{\gamma g}(z^s)$ and $x^s_f = \text{prox}_{\gamma f}(\text{refl}_{\gamma g}(z^s))$, for $s \in \{*, k\}$. Proximal and
reflection maps are nonexpansive, so we have the following simple bound:

\[
\max\{\|x^k_f - x^*\|, \|x^k_g - x^*\|\} \leq \|z^k - z^*\| \leq \|z^0 - z^*\|.
\]

Thus, \((x^j_f)_{j \geq 0}, (x^j_g)_{j \geq 0} \subseteq \overline{B}(x^*, \|z^0 - z^*\|)\). By the convexity of the closed ball, we also have \((\overline{x}^j_f)_{j \geq 0}, (\overline{x}^j_g)_{j \geq 0} \subseteq \overline{B}(x^*, \|z^0 - z^*\|)\).

**Corollary 2.5.1** (Ergodic convergence with single Lipschitz function). Let the notation be as in Theorem 2.5.1. Suppose that \(f\) (respectively \(g\)) is \(L\)-Lipschitz continuous on \(\overline{B}(x^*, \|z^0 - z^*\|)\), and let \(x^k = x^k_f\) (respectively \(x^k = x^k_g\)). Then the following convergence rate holds

\[
0 \leq f(\overline{x}^k) + g(\overline{x}^k) - f(x^*) - g(x^*) \leq \frac{1}{4\gamma \Lambda_k} \|z^0 - x^*\|^2 + \frac{2L\|z^0 - z^*\|}{\Lambda_k}.
\]

**Proof.** From Equation (2.5.1), we have \(\|\overline{x}_g^k - \overline{x}_f^k\| \leq (2/\Lambda_k)\|z^0 - z^*\|\). In addition, \((x^j_f)_{j \geq 0}, (x^j_g)_{j \geq 0} \subseteq \overline{B}(x^*, \|z^0 - z^*\|)\). Thus, it follows that

\[
0 \leq f(\overline{x}^k) + g(\overline{x}^k) - f(x^*) - g(x^*) \\
\leq f(\overline{x}_f^k) + g(\overline{x}_g^k) - f(x^*) - g(x^*) + L\|\overline{x}_f^k - \overline{x}_g^k\| \\
\overset{(2.5.1)}{\leq} f(\overline{x}_f^k) + g(\overline{x}_g^k) - f(x^*) - g(x^*) + \frac{2L\|z^0 - z^*\|}{\Lambda_k}.
\]

The upper bound follows from this equation and Theorem 2.5.1. \(\square\)

### 2.5.2 Nonergodic convergence rates

In this section, we prove the nonergodic convergence rate of the Algorithm 1 whenever \(\overline{\tau} := \inf_{j \geq 0} \tau_j > 0\). The proof uses Theorem 2.3.1 to bound the fundamental inequalities in Propositions 2.4.1 and 2.4.2.

**Theorem 2.5.2** (Nonergodic convergence of relaxed PRS). For all \(k \geq 0\), let \(\lambda_k \in (0, 1)\). Suppose that \(\overline{\tau} := \inf_{j \geq 0} \lambda_k (1 - \lambda_k) > 0\). Then we have the convergence rates:
1. In general, we have the bounds:

\[
\frac{\|z^0 - z^*\|\|z^* - x^*\|}{2\gamma \sqrt{\tau (k+1)}} \leq f(x^k_f) + g(x^k_g) - f(x^*) - g(x^*) \\
\leq \left( \frac{\|z^0 - z^\star\| + \|z^* - x^*\|\|z^0 - z^\star\|}{2\gamma \sqrt{\tau (k+1)}} \right)
\]

and \(|f(x^k_f) + g(x^k_g) - f(x^*) - g(x^*)| = o\left(1/\sqrt{k+1}\right)\).

2. If \(\mathcal{H} = \mathbb{R}\) and \(\lambda_k \equiv 1/2\), then for all \(k \geq 0\),

\[
\frac{\|z^0 - z^\star\|\|z^* - x^*\|}{\sqrt{2\gamma (k+1)}} \leq f(x^{k+1}_f) + g(x^{k+1}_g) - f(x^*) - g(x^*) \\
\leq \left( \frac{\|z^0 - z^\star\| + \|z^* - x^*\|\|z^0 - z^\star\|}{\sqrt{2\gamma (k+1)}} \right)
\]

and \(|f(x^{k+1}_f) + g(x^{k+1}_g) - f(x^*) - g(x^*)| = o\left(1/(k+1)\right)\).

**Proof.** We prove Part 1 first. For all \(\lambda \in [0, 1]\), let \(z^\lambda = (T_{PRS})_\lambda (z^k)\). Evaluate the upper inequality in Equation (2.4.6) at \(x = x^*\) to get

\[
4\gamma \lambda (f(x^k_f) + g(x^k_g) - f(x^*) - g(x^*)) \leq \|z^k - x^*\|^2 - \|z^\lambda - x^*\|^2 + \left(1 - \frac{1}{\lambda}\right) \|z^\lambda - z^k\|^2.
\]

Recall the following identity:

\[
\|z^k - x^*\|^2 - \|z^\lambda - x^*\|^2 - \|z^\lambda - z^k\|^2 = 2\langle z^\lambda - x^*, z^k - z^\lambda \rangle.
\]

By the triangle inequality, because \(\|z^\lambda - z^*\| \leq \|z^k - z^*\|\), and because \((\|z^j - z^*\|)_{j \geq 0}\) is monotonically nonincreasing (Corollary 2.3.1), it follows that

\[
\|z^\lambda - x^*\| \leq \|z^\lambda - z^*\| + \|z^* - x^*\| \leq \|z^0 - z^*\| + \|z^* - x^*\|. \quad (2.5.5)
\]
Thus, we have the bound:

\[
    f(x_f^k) + g(x_g^k) - f(x^*) - g(x^*)
    \leq \inf_{\lambda \in [0,1]} \frac{1}{4\gamma \lambda} \left( 2\langle z_{\lambda} - x^*, z^k - z_{\lambda} \rangle + 2 \left( 1 - \frac{1}{2\lambda} \right) \| z_{\lambda} - z^k \|^2 \right)
    \leq \frac{1}{\gamma} \| z_{1/2} - x^* \| \| z^k - z_{1/2} \|
    \leq (2.5.5)
    \leq \frac{1}{\gamma} \left( \| z^0 - z^* \| + \| z^* - x^* \| \right) \| z^k - z_{1/2} \|
    \leq (2.3.6) \frac{\| z^0 - z^* \| + \| z^* - x^* \| \| z^0 - z^* \|}{2\gamma \sqrt{\tau(k + 1)}}.
\]

The lower bound follows from the identity \( x_f^k - x_g^k = (1/2\lambda_k)(z^k - z^{k+1}) \) and the fundamental lower inequality in Equation (2.4.7):

\[
    f(x_f^k) + g(x_g^k) - f(x^*) - g(x^*) \geq \frac{1}{2\gamma \lambda_k} \langle z^k - z^{k+1}, z^* - x^* \rangle \geq -\frac{\| z^{k+1} - z^k \| \| z^* - x^* \|}{2\gamma \lambda_k}
    \geq (2.3.6) -\frac{\| z^0 - z^* \| \| z^* - x^* \|}{2\gamma \sqrt{\tau(k + 1)}}.
\]

Finally, the \( o(1/\sqrt{k + 1}) \) convergence rate follows from Equations (2.5.6) and (2.5.7) combined with Corollary 2.3.1 because each upper bound is of the form (bounded quantity) \( \times \sqrt{\text{FPR}} \), and \( \sqrt{\text{FPR}} \) has rate \( o(1/\sqrt{k + 1}) \).

Part 2 follows by the same analysis but uses Theorem 2.3.3 to estimate the FPR convergence rate.

Whenever \( f \) or \( g \) is Lipschitz, we can compute the convergence rate of \( f + g \) evaluated at the same point. The following theorem is analogous to Corollary 2.5.1 in the ergodic case. The proof essentially follows by combining the nonergodic convergence rate in Theorem 2.5.2 with the convergence rate of \( \| x_f^k - x_g^k \| = (1/\lambda_k)\| z^{k+1} - z^k \| \) deduced in Corollary 2.3.1.

**Corollary 2.5.2** (Nonergodic convergence with Lipschitz assumption). Let the notation be as in Theorem 2.5.2. Suppose that \( f \) (respectively \( g \)) is \( L \)-Lipschitz continuous on
and let $x_k = x_k^g$ (respectively $x_k = x_k^f$). Then we have the convergence rates of the nonnegative term:

1. In general, we have the bounds:

$$0 \leq f(x^k) + g(x^k) - f(x^*) - g(x^*) \leq \frac{\|z^0 - z^*\| + \|z^* - x^*\| + \gamma L \|z^0 - z^*\|}{2\gamma \sqrt{\tau(k + 1)}}$$

and $f(x^k) + g(x^k) - f(x^*) - g(x^*) = o\left(1/\sqrt{k + 1}\right)$.

2. If $\mathcal{H} = \mathbb{R}$ and $\lambda_k \equiv 1/2$, then for all $k \geq 0$,

$$0 \leq f(x^{k+1}) + g(x^{k+1}) - f(x^*) - g(x^*) \leq \frac{\|z^0 - z^*\| + \|z^* - x^*\| + \gamma L \|z^0 - z^*\|}{2\gamma \sqrt{(k + 1)}}$$

and $f(x^{k+1}) + g(x^{k+1}) - f(x^*) - g(x^*) = o\left(1/(k + 1)\right)$.

**Proof.** We prove Part 1 first. First recall that $\|x_k^g - x_k^f\| = (1/(2\lambda_k))\|z^{k+1} - z^k\|$. In addition, $(x_j^f)_{j \geq 0}, (x_j^g)_{j \geq 0} \subseteq \overline{B(x^*, \|z^0 - z^*\|)}$ (See Section (2.5.1)). Thus, it follows that

$$f(x^k) + g(x^k) - f(x^*) - g(x^*) \leq f(x_k^f) + g(x_k^g) - f(x^*) - g(x^*) + L\|x_k^f - x_k^g\|$$

$$= f(x_k^f) + g(x_k^g) - f(x^*) - g(x^*) + \frac{L\|z^{k+1} - z^k\|}{2\lambda_k} \quad (2.5.8)$$

$$\leq f(x_k^f) + g(x_k^g) - f(x^*) - g(x^*) + \frac{\gamma L \|z^0 - z^*\|}{2\gamma \sqrt{\tau(k + 1)}} \quad (2.5.9)$$

Therefore, the upper bound follows from Theorem 2.5.2 and Equation (2.5.9). In addition, the $o(1/\sqrt{k + 1})$ bound follows from Theorem 2.5.2 combined with Equation (2.5.8) and Corollary 2.3.1 because each upper bound is of the form (bounded quantity) $\times \sqrt{\text{FPR}}$, and $\sqrt{\text{FPR}}$ has rate $o(1/\sqrt{k + 1})$.

Part 2 follows by the same analysis, but uses Theorem 2.3.3 to estimate the FPR convergence rate.
2.6 Optimal FPR rate and arbitrarily slow convergence

In this section, we provide two examples where the DRS algorithm converges slowly. Both examples are a special cases of the following example, which originally appeared in [6, Section 7].

**Example 2.6.1 (DRS applied to two subspaces).** Let \( H = \ell_2^2(N) = \{(z_j)_{j \geq 0} | \forall j \in N, z_j \in \mathbb{R}^2, \sum_{i=0}^{\infty} \|z^i\|^2 < \infty\} \). Let \( R_\theta \) denote counterclockwise rotation in \( \mathbb{R}^2 \) by \( \theta \) degrees.

Let \( e_0 := (1, 0) \) denote the standard unit vector, and let \( e_\theta := R_\theta e_0 \). Suppose that \((\theta_j)_{j \geq 0}\) is a sequence of angles in \((0, \pi/2] \) such that \( \theta_i \to 0 \) as \( i \to \infty \). We define two subspaces:

\[
U := \bigoplus_{i=0}^{\infty} Re_0 \quad \text{and} \quad V := \bigoplus_{i=0}^{\infty} R e_{\theta_i},
\]

where \( Re_0 = \{\alpha e_0 : \alpha \in \mathbb{R}\} \) and \( Re_{\theta_i} = \{\alpha e_{\theta_i} : \alpha \in \mathbb{R}\} \). See Figure 2.2 for an illustration.

Note that [6, Section 7] shows the projection identities

\[
(P_V)_i = \begin{bmatrix}
\cos^2(\theta_i) & \sin(\theta_i) \cos(\theta_i) \\
\sin(\theta_i) \cos(\theta_i) & \sin^2(\theta_i)
\end{bmatrix}
\quad \text{and} \quad
(P_U)_i = \begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix},
\]

the DRS operator identity

\[
T := (T_{PRS})_{1/2} = c_0 R_{\theta_0} \oplus c_1 R_{\theta_1} \oplus \cdots
\]

and that \((z^i)_{i \geq 0}\) converges in norm to \( z^* = 0 \) for any initial point \( z^0 \).

2.6.1 Optimal FPR rates

The following theorem shows that the FPR estimates derived in Corollary 2.3.1 are essentially optimal. We note that this is the first optimality result for the FPR of the DRS iteration in the case of variational problems.

**Theorem 2.6.1 (Lower FPR complexity of DRS).** There exists a Hilbert space \( H \) and two closed subspaces \( U \) and \( V \) with zero intersection, \( U \cap V = \{0\} \), such that for every
\[ \theta_1 \quad \oplus \quad \theta_2 \quad \oplus \quad \ldots \]

**Figure 2.2:** Illustration of Example 2.6.1. Each pair of lines represents a 2-dimensional component of \( U \cup V \). The angles \( \theta_k \) are converging to 0.

\( \alpha > 1/2 \), there exists \( z^0 \in \mathcal{H} \) such that if \( (z^j)_{j \geq 0} \) is generated by \( T = (T_{PRS})^{1/2} \) applied to \( f = \iota_V \) and \( g = \iota_U \), then for all \( k \geq 1 \), we have the bound:

\[
\| Tz^k - z^k \|^2 \geq \frac{1}{(k + 1)^{2\alpha}}.
\]

**Proof.** We assume the setting of Example 2.6.1. For all \( i \geq 0 \) set \( c_i = \left( \frac{i}{i + 1} \right)^{1/2} \), and let \( w^0 = (w^0_j)_{j \geq 0} \in \mathcal{H} \), where each \( w^0_i \in \mathbb{R}^2 \) satisfies \( \|w^0_i\| = \sqrt{2\alpha e(i + 1)^{-1+2\alpha}/2} \). Then for all \( k \geq 1 \),

\[
\| T^k w^0 \|^2 = \sum_{i=0}^{\infty} c_i^{2k} \|w^0_i\|^2 \geq \sum_{i=k}^{\infty} \left( \frac{i}{i + 1} \right)^k \frac{2\alpha e}{(i + 1)^{1+2\alpha}} \geq \frac{1}{(k + 1)^{2\alpha}}
\]

where we have used the bound \( (i/(i + 1))^k \geq e^{-1} \) for \( i \geq k \) and the lower integral approximation of the sum.

Now we will show that \( w^0 \) is in the range of \( I - T \). Indeed, for all \( i \geq 1 \) each block of \( I - T \) is of the form

\[
I_{\mathbb{R}^2} - \cos(\theta_i) R_{\theta_i} = \begin{bmatrix} \sin^2(\theta_i) & \sin(\theta_i) \cos(\theta_i) \\ -\sin(\theta_i) \cos(\theta_i) & \sin^2(\theta_i) \end{bmatrix} = \begin{bmatrix} \frac{1}{i+1} & \frac{\sqrt{i}}{i+1} \\ -\frac{\sqrt{i}}{i+1} & \frac{1}{i+1} \end{bmatrix}. \tag{2.6.3}
\]

Therefore, the point \( z^0 = (\sqrt{2\alpha e}((1/(j+1)^\alpha, 0))_{j \geq 0} \in \mathcal{H} \) has image

\[
w^0 = (I - T)z^0 = \left( \sqrt{2\alpha e} \left( \frac{1}{(j+1)^{\alpha+1}}, \frac{-\sqrt{j}}{(j+1)^{\alpha+1}} \right) \right)_{j \geq 0}.
\]

In addition, for all \( i \geq 1 \), we have \( \|w^0_i\| = \sqrt{2\alpha e(i + 1)^{-1+2\alpha}/2} \), and the inequality follows. \( \square \)
Remark 2.6.1. The proof of Theorem 2.6.1 crucially relies on the strictness of inequality, $\alpha > 1/2$: if $\alpha = 1/2$, then $\|z^0\| = \infty$.

2.6.1.1 Notes on Theorem 2.6.1

With this new optimality result in hand, we can make the following list of optimal FPR rates, not to be confused with optimal rates in objective error, for a few standard splitting schemes:

**Proximal point algorithm (PPA):** For the general class of monotone operators, the counterexample furnished in [28, Remarque 4] shows that there exists a maximal monotone operator $A$ such that when iteration (2.3.1) is applied to the resolvent $J_{\gamma A}$, the rate $o(1/(k+1))$ is tight. In addition, if $A = \partial f$ for some closed, proper, and convex function $f$, then the FPR rate improves to $O(1/(k+1)^2)$ [28, Théorème 9]. We improve this result to $o(1/(k+1)^2)$ in Theorem 2.3.3. This result appears to be new and is optimal by [28, Remarque 6].

**Forward backward splitting (FBS):** The FBS method reduces to the proximal point algorithm when the differentiable (or single valued operator) term is trivial. Thus, for the general class of monotone operators, the $o(1/(k+1))$ FPR rate is optimal by [28, Remarque 4]. We improve this rate to $o(1/(k+1)^2)$ in Theorem 2.3.3. This result appears to be new, and is optimal by [28, Remarque 6].

**Douglas-Rachford splitting/ADMM:** Theorem 2.6.1 shows that the optimal FPR rate is $o(1/(k+1))$. Because the DRS iteration is equivalent to a proximal point iteration applied to a special monotone operator [60, Section 4], Theorem 2.6.1 provides an alternative counterexample to [28, Remarque 4]. In particular, Theorem 2.6.1 shows that, in general, there is no closed, proper, and convex function $f$ such that $(T_{PRS})_{1/2} = \text{prox}_{\gamma f}$. In the one dimensional case, we improve the FPR to $o(1/(k+1)^2)$ in Theorem 2.3.4.

**Miscellaneous methods:** By similar arguments we can deduce the tight FPR iteration complexity for the following methods, each of which at least has rate $o(1/(k+1))$.
by Theorem 2.3.1: Standard Gradient descent \( o(1/(k+1)^2) \): (the rate follows from Theorem 2.3.3. Optimality follows from the fact that PPA is equivalent to gradient descent on Moreau envelope \([11, \text{Proposition 12.29}] \) and \([28, \text{Remarque 4}] \); Forward-Douglas Rachford splitting \([30]\): \( o(1/(k+1)) \) (choose a trivial cocoercive operator and use Theorem 2.6.1); Chambolle and Pock’s primal-dual algorithm \([36]\) \( o(1/(k+1)) \): (reduce to DRS (\( \sigma = \tau = 1 \)) \([36, \text{Section 4.2}] \) and apply Theorem 2.6.1 using the transformation \( z^k = \text{primal}_k + \text{dual}_k \) \([36, \text{Equation (24)}] \) and the lower bound

\[
\|z^{k+1} - z^k\|^2 \leq 2\|\text{primal}_{k+1} - \text{primal}_k\|^2 + 2\|\text{dual}_{k+1} - \text{dual}_k\|^2;
\]

Vũ/Condat’s primal-dual algorithm \([112, 52]\) \( o(1/(k+1)) \): (reduces to Chambolle and Pock’s method \([36]\)).

Note that the rate established in Theorem 2.3.1 has broad applicability, and this list is hardly extensive. For PPA, FBS, and standard gradient descent, the FPR always has rate that is the square of the objective value convergence rate. We will see that the same is true for DRS in Theorem 2.7.2.

### 2.6.2 Arbitrarily slow convergence

In \([6, \text{Section 7}] \), the DRS setting in Example 2.6.1 is shown to converge in norm, but not linearly. We improve their result by showing that a proper choice of parameters yields arbitrarily slow convergence in norm.

The following technical lemma will help us construct a sequence that converges arbitrarily slowly. The idea of the proof follows directly from the proof of \([62, \text{Theorem 4.2}] \), which shows that the alternating projection algorithm can converge arbitrarily slowly.

**Lemma 2.6.1.** Suppose that \( h : \mathbb{R}_+ \to (0, 1) \) is a function that is monotonically decreasing to zero. Then there exists a monotonic sequence \((c_j)_{j \geq 0} \subseteq (0, 1) \) such that \( c_k \to 1^- \) as \( k \to \infty \) and an increasing sequence of integers \((n_j)_{j \geq 0} \subseteq \mathbb{N} \cup \{0\} \) such that for all \( k \geq 0 \),

\[
\frac{c_{n_k}^{k+1}}{n_k + 1} > h(k + 1)e^{-1}.
\]  

(2.6.4)
Proof. Let \( h_2 \) be the inverse of the strictly increasing function \((1/h) - 1\), let \([x]\) denote the integer part of \(x\), and for all \(k \geq 0\) let
\[
c_k = \frac{h_2(k + 1)}{1 + h_2(k + 1)}.
\]
(2.6.5)

Then \((c_j)_{j \geq 0}\) is monotonic and \(c_k \to 1^-\). For all \(x \geq 0\), we have \(h_2^{-1}(x) = 1/h(x) - 1 \leq [1/h(x)]\), thus, \(x \leq h_2([1/h(x)])\). Now, choose \(n_k \geq 0\) such that \(n_k + 1 = [1/h(k + 1)]\).

Therefore,
\[
\frac{c_{n_k}^{k+1}}{n_k + 1} \geq h(k + 1) \left( \frac{k + 1}{1 + (k + 1)} \right)^{k+1} \geq h(k + 1)e^{-1}.
\]

\(\square\)

**Theorem 2.6.2** (Arbitrarily slow convergence of DRS). There is a point \(z_0 \in \ell^2_2(\mathbb{N})\), such that for every function \(h : \mathbb{R}_+ \to (0, 1)\) that strictly decreases to zero, there exists two closed subspaces \(U\) and \(V\) with zero intersection, \(U \cap V = \{0\}\), such that the relaxed PRS sequence \((z^j)_{j \geq 0}\) generated with the functions \(f = \iota_V\) and \(g = \iota_U\) and relaxation parameters \(\lambda_k \equiv 1/2\) satisfies the bound
\[
\|z^k - z^*\| \geq e^{-1}h(k)
\]
but \((\|z^j - z^*\|)_{j \geq 0}\) converges to 0.

Proof. We assume the setting of Example 2.6.1. Suppose that \(z^0 = (z^0_j)_{j \geq 0}\), where for all \(k \geq 0\), \(z^0_k \in \mathbb{R}^2\), and \(\|z^0_k\| = 1/(k + 1)\). Then it follows that \(\|z^0\|_H^2 = \sum_{i=0}^{\infty} 1/(k+1)^2 < \infty\) and so \(z^0 \in \mathcal{H}\). Thus, for all \(k, n \geq 0\)
\[
\|T^{k+1}z^0\| \geq c_{n}^{k+1}\|z^0_n\| = \frac{1}{n + 1}c_{n}^{k+1}.
\]
(2.6.6)

Therefore, we can achieve arbitrarily slow convergence by picking \((c_j)_{j \geq 0}\), and a subsequence \((n_j)_{j \geq 0} \subseteq \mathbb{N}\) using Lemma 2.6.1. \(\square\)
2.7 Optimal objective rates

In this section we construct four examples that show the nonergodic and ergodic convergence rates in Corollary 2.5.2 and Theorem 2.5.1 are optimal up to constant factors. In particular, we provide examples of optimal ergodic convergence in the minimization case and in the feasibility case, where no objective is driving the minimization.

2.7.1 Ergodic convergence of feasibility problems

Proposition 2.7.1. The ergodic feasibility convergence rate in Equation (2.5.1) is optimal up to a factor of two.

Proof. Figure 2.3 shows Algorithm 1 with $\lambda_k = 1$ for all $k \geq 0$ (i.e. PRS) applied to the functions $f = \iota_{\{(x_1,x_2)\in\mathbb{R}^2|x_1=0\}}$ and $g = \iota_{\{(x_1,x_2)\in\mathbb{R}^2|x_2=0\}}$ with the initial iterate $z^0 = (1, 1) \in \mathbb{R}^2$. Because $T_{PRS} = -I_H$, it is easy to see that the only fixed point of $T_{PRS}$ is $z^* = (0, 0)$. In addition, the following identities are satisfied:

$$x^k_g = \begin{cases} (1, 0) & \text{even } k; \\ (-1, 0) & \text{odd } k. \end{cases}$$

$$z^k = \begin{cases} (1, 1) & \text{even } k; \\ (-1, -1) & \text{odd } k. \end{cases}$$

$$x^k_f = \begin{cases} (0, -1) & \text{even } k; \\ (0, 1) & \text{odd } k. \end{cases}$$

Thus, the PRS algorithm oscillates around the solution $x^* = (0, 0)$. However, note that the averaged iterates satisfy:

$$\bar{x}^k_g = \begin{cases} \left(\frac{1}{k+1}, 0\right) & \text{even } k; \\ (0, 0) & \text{odd } k. \end{cases}$$

$$\bar{x}^k_f = \begin{cases} \left(0, \frac{-1}{k+1}\right) & \text{even } k; \\ (0, 0) & \text{odd } k. \end{cases}$$

It follows that $\|\bar{x}^k_g - \bar{x}^k_f\| = (1/(k+1))\| (1, -1)\| = (1/(k+1))\| z^0 - z^* \|$, for all $k \geq 0$. 

2.7.2 Ergodic convergence of minimization problems

In this section, we will construct an example where the ergodic rates of convergence in Section 2.5.1 are optimal up to constant factors. In addition, the example we construct
only converges in the ergodic sense and diverges otherwise. Throughout this section, we let \( \gamma = 1 \) and \( \lambda_k \equiv 1 \), we work in the Hilbert space \( \mathcal{H} = \mathbb{R} \), and we consider the following objective functions: for all \( x \in \mathbb{R} \), let

\[
g(x) = 0, \quad \text{and} \quad f(x) = |x|. \quad (2.7.1)
\]

Recall that for all \( x \in \mathbb{R} \)

\[
\text{prox}_g(x) = x, \quad \text{and} \quad \text{prox}_f(x) = \max (|x| - 1, 0) \text{ sign}(x). \quad (2.7.2)
\]

The following lemma characterizes the minimizer of \( f + g \) and the fixed points of \( T_{\text{PRS}} \). The proof is simple so we omit it.

**Lemma 2.7.1.** The minimizer of \( f + g \) is unique and equal to \( 0 \in \mathbb{R} \). Furthermore, \( 0 \) is the unique fixed point of \( T_{\text{PRS}} \).

Because of Lemma 2.7.1, we will use the notation:

\[
z^* = 0 \quad \text{and} \quad x^* = 0. \quad (2.7.3)
\]
We are ready to prove our main optimality result.

**Proposition 2.7.2** (Optimality of ergodic convergence rates). Suppose that \( z^0 = 2 - \varepsilon \) for some \( \varepsilon \in (0, 1) \). Then the PRS algorithm applied to \( f \) and \( g \) with initial point \( z^0 \) does not converge.

Furthermore, as \( \varepsilon \) goes to 0, the ergodic objective convergence rate in Theorem 2.5.1 is tight, and the ergodic objective convergence rate in Corollary 2.5.1 is tight up to a factor of \( 5/2 \). In addition, the feasibility convergence rate of Theorem 2.5.1 is tight up to a factor of 4.

**Proof.** We will now compute the sequences \((z^i)_{i \geq 0}\), \((x^i_g)_{i \geq 0}\), and \((x^i_f)_{i \geq 0}\). We proceed by induction: First \( x^0_g = \text{prox}_g(z^0) = z^0 \) and \( x^0_f = \text{prox}_f(2x^0_g - z^0) = \max(\{|z^0| - 1, 0\} \text{sign}(z^0) = 1 - \varepsilon) \). Thus, it follows that \( z^1 = z^0 + 2(x^1_f - x^0_t) = 2 - \varepsilon + 2(1 - \varepsilon - (2 - \varepsilon)) = z^0 = -\varepsilon \). Similarly, \( x^1_g = z^1 = -\varepsilon \). Finally, \( x^1_f = \max(\varepsilon - 1, 0) \text{sign}(-\varepsilon) = 0 \) and \( z^2 = z^1 + 2(x^1_f - x^0_g) = z^1 + 2(\varepsilon) = \varepsilon \). Thus, by induction we have the following identities: For all \( k \geq 1 \),

\[
\begin{align*}
z^k &= (-1)^k \varepsilon, \\
x^k_g &= (-1)^k \varepsilon, \\
x^k_f &= 0.
\end{align*}
\]

Notice that \((z^i)_{i \geq 0}\) and \((x^i_g)_{i \geq 0}\) do not converge, but they oscillate around the fixed point of \( T_{\text{PRS}} \).

We will now compute the ergodic iterates:

\[
\bar{x}^k_g = \frac{1}{k+1} \sum_{i=0}^{k} x^i_g = \begin{cases} \\
\frac{2-\varepsilon}{k+1} & \text{if } k \text{ is even}; \\
\frac{2-2\varepsilon}{k+1} & \text{otherwise}. \\
\end{cases}
\]

\[
\bar{x}^k_f = \frac{1}{k+1} \sum_{i=0}^{k} x^i_f = \begin{cases} \\
\frac{1-\varepsilon}{k+1}; \\
\frac{2-2\varepsilon}{k+1} & \text{otherwise}. \\
\end{cases}
\]

(2.7.4)

(2.7.5)

Let us use these formulas to compute the objective values:

\[
f(\bar{x}^k_f) + g(\bar{x}^k_g) - f(0) - g(0) = \begin{cases} \\
\frac{1-\varepsilon}{k+1}; \\
\frac{2-2\varepsilon}{k+1} & \text{if } k \text{ is even}; \\
\frac{2-2\varepsilon}{k+1} & \text{otherwise}. \\
\end{cases}
\]

(2.7.5)

(2.7.6)
We will now compare the theoretical bounds from Theorem 2.5.1 and Corollary 2.5.1 with the rates we observed in Equation (2.7.6). Theorem 2.5.1 bounds the objective error at $x^k_f$ by

$$\frac{|z^0 - x^*|^2}{4(k+1)} = \frac{4 - 4\varepsilon}{4(k+1)} + \frac{\varepsilon^2}{4(k+1)} = \frac{1 - \varepsilon}{k+1} + \frac{\varepsilon^2}{4(k+1)}.$$  \hspace{1cm} (2.7.7)

By taking $\varepsilon$ to 0, we see that this bound is tight.

Because $f$ is 1-Lipschitz continuous, Corollary 2.5.1 bounds the objective error at $x^k_g$ with

$$\frac{|z^0 - x^*|^2}{4(k+1)} + \frac{2|z^0 - z^*|}{(k+1)} = \frac{1 - \varepsilon}{k+1} + \frac{\varepsilon^2}{4(k+1)} + 2\frac{2 - \varepsilon}{k+1} = \frac{5 - 3\varepsilon}{k+1} + \frac{\varepsilon^2}{4(k+1)}.$$ \hspace{1cm} (2.7.8)

As we take $\varepsilon$ to 0, we see that this bound it tight up to a factor of 5/2.

Finally, consider the feasibility convergence rate:

$$|x^k_g - x^k_f| = \begin{cases} \frac{1}{k+1} & \text{if } k \text{ is even;} \\ \frac{1 - \varepsilon}{k+1} & \text{otherwise}. \end{cases} \hspace{1cm} (2.7.9)$$

Theorem 2.5.1 predicts the following upper bound for Equation (2.7.9):

$$\frac{2|z^0 - z^*|}{k+1} = 2\frac{2 - \varepsilon}{k+1} = \frac{4 - 2\varepsilon}{k+1}.$$  \hspace{1cm} (2.7.10)

By taking $\varepsilon$ to 0, we see that this bound is tight up to a factor of 4.

2.7.3 Optimal nonergodic objective rates

Our aim in this section is to show that if $\lambda_k \equiv 1/2$, then the non-ergodic convergence rate of $o(1/\sqrt{k+1})$ in Corollary 2.5.2 is essentially tight. In particular, for every $\alpha > 1/2$, we provide examples of $f$ and $g$ such that $f$ is 1-Lipschitz and

$$f(x^k_g) + g(x^k_g) - f(x^*) - g(x^*) = \Omega \left( \frac{1}{(k+1)^\alpha} \right).$$

Throughout this section, we will be working with the proximal operator of a distance functions.
Proposition 2.7.3. Let $C$ be a closed, convex subset of $\mathcal{H}$ and let $d_C(x) = \min_{y \in C} \|x - y\|$. Then $d_C(x)$ is 1-Lipschitz and for all $x \in \mathcal{H}$

$$\text{prox}_{\gamma d_C}(x) = \theta P_C(x) + (1 - \theta)x$$

where

$$\theta = \begin{cases} \frac{\gamma}{d_C(x)} & \text{if } \gamma \leq d_C(x); \\ 1 & \text{otherwise.} \end{cases} \quad (2.7.11)$$

Proof. Follows directly from the formula for the subgradient of $d_C$ [11, Example 16.49].

Proposition 2.7.3 says that $\text{prox}_{\gamma d_C}(x)$ reduces to a projection map whenever $x$ is close enough to $C$. Proposition 2.7.4 constructs a family of examples such that if $\gamma$ is chosen large enough, then DRS does not distinguish between indicator functions and distance functions.

Proposition 2.7.4. Suppose that $V$ and $U$ are linear subspaces of $\mathcal{H}$ and $U \cap V = \{0\}$. If $\gamma \geq \|z^0\|$ and $\lambda_k = 1/2$ for all $k \geq 0$, then Algorithm 1 applied to the either pair of objective functions $(f = \iota_V, g = \iota_U)$ and $(f = d_V, g = \iota_U)$ produces the same sequence $(z^j)_{j \geq 0}$

Proof. Let $(z^j)_{j \geq 0}$ be the sequence generated by the functions $(f = \iota_V, g = \iota_U)$. Observe that $x^* = 0$ is a minimizer of both functions pairs and $z^* = 0$ is a fixed point of $(T_{PRS})_{1/2}$. In particular, we set $\tilde{\nabla}_{t_V}(x^*) = P_V(\text{refl}_g(z^*)) - x^* = 0$. Therefore, we just need to show that $\text{prox}_{\gamma d_V}(\text{refl}_g(z^k)) = P_V(\text{refl}_g(z^k))$ for all $k \geq 0$. Note that by definition, $x^k_{i_V} = P_V(\text{refl}_g(z^k))$ and $\tilde{\nabla}_{t_V}(x^k_{i_V}) = \text{refl}_g(z^k) - P_V(\text{refl}_g(z^k)) \in \partial\iota_V(x^k_{i_V})$. In view of Proposition 2.7.3, the identity will follow if

$$\gamma \geq d_V(\text{refl}_g(z^k)) = \|\text{refl}_g(z^k) - P_V(\text{refl}_g(z^k))\| = \|\tilde{\nabla}_{t_V}(x^k_{i_V})\| = \|\tilde{\nabla}_{t_V}(x^k_{i_V}) - \tilde{\nabla}_{t_V}(x^*)\|.$$

However, this is always the case because

$$\|\tilde{\nabla}_{t_V}(x^k_{i_V}) - \tilde{\nabla}_{t_V}(x^*)\|^2 + \|x^k_{i_V} - x^*\|^2 \overset{2.1.16}{\leq} \|\text{refl}_g(z^k) - \text{refl}_g(z^*)\|^2 \leq \|z^k - z^*\|^2 \leq \|z^0 - z^*\|^2 \leq \|z^0\|^2 \leq \gamma^2.$$
\textbf{Theorem 2.7.1.} Assume the notation of Theorem 2.6.1. Then for all $\alpha > 1/2$, there exists a point $z^0 \in \mathcal{H}$ such that if $\gamma \geq \|z^0\|$ and $(z^j)_{j \geq 0}$ is generated by DRS applied to the functions $(f = d_V, g = \iota_U)$, then $d_V(x^*) = 0$ and
\begin{equation}
d_V(x_g^k) = \Omega \left( \frac{1}{(k+1)^\alpha} \right). \tag{2.7.12}\end{equation}

\textit{Proof.} Let $z^0 = ((1/(j+1)\alpha, 0))_{j \geq 0} \in \mathcal{H}$. Now, choose $\gamma^2 \geq \|z^0\|^2 = \sum_{i=0}^{\infty} 1/(i+1)^{2\alpha}$. Define $w^0 \in \mathcal{H}$ using Equation (2.6.3):
\begin{align*}
w^0 &= (I - T)z^0 = \left( \frac{1}{(j+1)^\alpha} \left( \frac{1}{j+1}, -\sqrt{j} \right) \right)_{j \geq 0}.
\end{align*}
Then $\|w^0\| = 1/(1 + i)^{(1+2\alpha)/2}$.

Now we will calculate $d_V(x_g^k) = \|P_Vx_g^k - x_g^k\|$. First, recall that $T^k = c_0 R_{k\theta_0} \oplus c_1 R_{k\theta_1} \oplus \cdots$, where
\begin{equation*}
R_\theta = \begin{bmatrix}
\cos(\theta) & -\sin(\theta) \\
\sin(\theta) & \cos(\theta)
\end{bmatrix}.
\end{equation*}
Thus,
\begin{align*}
x_g^k := P_V(z^k) &= \begin{bmatrix}
1 \\
0
\end{bmatrix} c_j^k \left( R_{k\theta_j} \left( \frac{1}{(j+1)^\alpha}, 0 \right) \right)_{j \geq 0} \\
&= \begin{bmatrix}
1 \\
0
\end{bmatrix} c_j^k \frac{1}{(j+1)^\alpha} \left( \cos(k\theta_j), \sin(k\theta_j) \right)_{j \geq 0} \\
&= \left( c_j^k \frac{\cos(k\theta_j)}{(j+1)^\alpha} (1, 0) \right)_{j \geq 0}.
\end{align*}
Furthermore, from the identity
\begin{equation*}
(P_V)_i = \begin{bmatrix}
\cos^2(\theta_i) & \sin(\theta_i) \cos(\theta_i) \\
\sin(\theta_i) \cos(\theta_i) & \sin^2(\theta_i)
\end{bmatrix} = \begin{bmatrix}
\frac{i}{i+1} & \frac{\sqrt{i}}{i+1} \\
\frac{\sqrt{i}}{i+1} & \frac{1}{i+1}
\end{bmatrix},
\end{equation*}
we have
\begin{align*}
P_Vx_g^k &= \left( c_j^k \frac{\cos(k\theta_j)}{(j+1)^\alpha} \left( \frac{j}{j+1}, \sqrt{j} \right) \right)_{j \geq 0}.
\end{align*}
Thus, the difference has the following form:

\[
 x^k_g - P_V x^k_g = \left( c^k_j \cos \left( \frac{k \theta_j}{j} \right) \left( \frac{1}{j+1}, -\frac{\sqrt{j}}{j+1} \right) \right)_{j \geq 0}.
\]

Now we derive the lower bound:

\[
 d_V(x^k_g)^2 = \| x^k_g - P_V x^k_g \|^2 = \sum_{i=0}^{\infty} c_i^{2k} \cos^2 \left( k \cos^{-1} \left( \sqrt{\frac{i}{i+1}} \right) \right) \left( \frac{1}{i+1}, -\frac{\sqrt{i}}{i+1} \right)^2 |
\]

\[
 \geq \frac{1}{e} \sum_{i=k}^{\infty} \cos^2 \left( k \cos^{-1} \left( \sqrt{\frac{i}{i+1}} \right) \right) \left( \frac{1}{i+1}, -\frac{\sqrt{i}}{i+1} \right)^2. \tag{2.7.13}
\]

The next several lemmas will focus on estimating the order of the sum in Equation (2.7.13). After which, Theorem 2.7.1 will follow from Equation (2.7.13) and Lemma 2.7.4, below. This completes the proof of Theorem 2.7.1.

Lemma 2.7.2. Let \( h : \mathbb{R}_+ \to \mathbb{R}_+ \) be a continuously differentiable function such that \( h \in L_1(\mathbb{R}_+) \) and \( \sum_{i=1}^{\infty} h(i) < \infty \). Then for all positive integers \( k \),

\[
 \left| \sum_{i=k}^{\infty} h(i) - \sum_{i=k}^{\infty} h(i) \right| \leq \sum_{i=k}^{\infty} \max_{y \in [i,i+1]} |h'(y)|.
\]

Proof. We just apply the Mean Value Theorem and combine the integral with the sum

\[
 \left| \int_{i}^{\infty} h(y)dy - \sum_{i=k}^{\infty} h(i) \right| \leq \sum_{i=k}^{\infty} \left| \int_{i}^{i+1} (h(y) - h(i))dy \right| \leq \sum_{i=k}^{\infty} \int_{i}^{i+1} |h(y) - h(i)|dy
\]

\[
 \leq \sum_{i=k}^{\infty} \max_{y \in [i,i+1]} |h'(y)|. \]

The following Lemma will quantify the deviation of integral from the sum.

Lemma 2.7.3. The following approximation bound holds:

\[
 \left| \sum_{i=k}^{\infty} \cos^2 \left( k \cos^{-1} \left( \sqrt{\frac{i}{i+1}} \right) \right) \left( \frac{1}{i+1}, -\frac{\sqrt{i}}{i+1} \right)^2 - \int_{i}^{\infty} \cos^2 \left( k \cos^{-1} \left( \sqrt{\frac{y}{y+1}} \right) \right) \frac{dy}{(y+1)^{2\alpha+1}} \right| = O \left( \frac{1}{(k+1)^{2\alpha+1/2}} \right). \tag{2.7.14}
\]
Proof. We will use Lemma 2.7.2 with
\[ h(y) = \frac{\cos^2 \left( k \cos^{-1} \left( \sqrt{\frac{y}{y+1}} \right) \right)}{(y+1)^{2\alpha+1}}. \]
to deduce an upper bound on the absolute value. Indeed,
\[ |h'(y)| = \left| \frac{k \sin \left( k \cos^{-1} \left( \sqrt{\frac{y}{y+1}} \right) \right) \cos \left( k \cos^{-1} \left( \sqrt{\frac{y}{y+1}} \right) \right)}{\sqrt{y}(y+1)(y+1)^{2\alpha+1}} - \frac{\cos^2 \left( k \cos^{-1} \left( \sqrt{\frac{y}{y+1}} \right) \right)}{(y+1)^{2\alpha+2}} \right|. \]
\[ = O \left( \frac{k}{(y+1)^{2\alpha+1+3/2}} + \frac{1}{(y+1)^{2\alpha+2}} \right). \]

Therefore, we can bound Equation (2.7.14) by the following sum:
\[ \sum_{i=k}^{\infty} \max_{y \in [i,i+1]} |h'(y)| = O \left( \frac{k}{(k+1)^{2\alpha+3/2}} + \frac{1}{(k+1)^{2\alpha+1}} \right) = O \left( \frac{1}{(k+1)^{2\alpha+1/2}} \right). \]

In the following Lemma, we estimate the order of the oscillatory integral approximation to the sum in Equation (2.7.13). The proof follows by a change of variables and an integration by parts.

**Lemma 2.7.4.** The following bound holds:
\[ \sum_{i=k}^{\infty} \cos^2 \left( k \cos^{-1} \left( \sqrt{\frac{i}{i+1}} \right) \right) dy = \Omega \left( \frac{1}{(k+1)^{2\alpha}} \right). \]  

(2.7.15)

Proof. Fix \( k \geq 1 \). We first perform a change of variables \( u = \cos^{-1}(\sqrt{y/(y+1)}) \) on the integral approximation of the sum:
\[ \int_{k}^{\infty} \cos^2 \left( k \cos^{-1} \left( \sqrt{\frac{y}{y+1}} \right) \right) dy = 2 \int_{0}^{\cos^{-1}(\sqrt{k/(k+1)})} \cos^2(ku) \cos(u) \sin^{4\alpha-1}(u) du. \]  

(2.7.16)

We will show that the right hand side of Equation (2.7.16) is of order \( \Omega \left( 1/(k+1)^{2\alpha} \right) \).
Then Equation (2.7.15) will follow by Lemma 2.7.3.
Let $\rho := \cos^{-1}(\sqrt{k/(k+1)})$. We have

\[
2 \int_0^\rho \cos^2(ku) \cos(u) \sin^{4\alpha - 1}(u) du = \int_0^\rho (1 + \cos(2ku)) \cos(u) \sin^{4\alpha - 1}(u) du
\]

\[
= p_1 + p_2 + p_3
\]

where

\[
p_1 = \int_0^\rho 1 \cdot \cos(u) \sin^{4\alpha - 1}(u) du = \frac{1}{4\alpha} \sin^{4\alpha}(\rho);
\]

\[
p_2 = \frac{1}{2k} \sin(2k\rho) \cos(\rho) \sin^{4\alpha - 1}(\rho);
\]

\[
p_3 = -\frac{1}{2k} \int_0^\rho \sin(2ku) d(cos(u) \sin^{4\alpha - 1}(u));
\]

and we have applied integration by parts for \( \int_0^\rho \cos(2ku) \cos(u) \sin^{4\alpha - 1}(u) du = p_2 + p_3. \)

Because \( \sin(\cos^{-1}(x)) = \sqrt{1 - x^2} \), for all \( \eta > 0 \), we get

\[
\sin^n(\rho) = \sin^n \cos^{-1}\left(\sqrt{k/(k+1)}\right) = \frac{1}{(k+1)^{\eta/2}}.
\]

In addition, we have \( \cos(\rho) = \cos \cos^{-1}\left(\sqrt{k/(k+1)}\right) = \sqrt{k/(k+1)} \) and the trivial bounds \( |\sin(2k\rho)| \leq 1 \) and \( |\sin(2ku)| \leq 1. \)

Therefore, the following bounds hold:

\[
p_1 = \frac{1}{4\alpha(k+1)^{2\alpha}} \quad \text{and} \quad |p_2| \leq \frac{\sqrt{k/(k+1)}}{2k(k+1)^{2\alpha - 1/2}} = O\left(\frac{1}{(k+1)^{2\alpha + 1/2}}\right).
\]

In addition, for \( p_3 \), we have \( d(\cos(u) \sin^{4\alpha - 1}(u)) = \sin^{4\alpha - 2}(u)((4\alpha - 1) \cos(u) - \sin^2(u)) du \).

Furthermore, for \( u \in [0, \rho] \) and \( \alpha > 1/2 \), we have \( \sin^{4\alpha - 2}(u) \in [0, 1/(k+1)^{2\alpha - 1}] \) and the following lower bound: \( (4\alpha - 1) \cos(u) - \sin^2(u) \geq (4\alpha - 1) \cos(\rho) - \sin^2(\rho) = (4\alpha - 1)\sqrt{k/(k+1)} - 1/(k+1) > 0 \) as long as \( k \geq 1. \) Therefore, we have \( \sin^{4\alpha - 2}(u)((4\alpha - 1) \cos(u) - \sin^2(u)) \geq 0 \) for all \( u \in [0, \rho] \) and, thus,

\[
|p_3| \leq \frac{1}{2k} \cos(\rho) \sin^{4\alpha - 1}(\rho) = \frac{\sqrt{k/(k+1)}}{2k(k+1)^{2\alpha - 1/2}} = O\left(\frac{1}{(k+1)^{2\alpha + 1/2}}\right).
\]

Therefore, \( p_1 + p_2 + p_3 \geq p_1 - |p_2| - |p_3| = \Omega\left( (k+1)^{-2\alpha} \right). \) \hfill \Box

We deduce the following theorem from the sum estimation in Lemma 2.7.4:

\[
\]
Theorem 2.7.2 (Lower complexity of DRS). There exists closed, proper, and convex functions \(f, g : \mathcal{H} \to (-\infty, \infty]\) such that \(f\) is 1-Lipschitz and for every \(\alpha > 1/2\), there is a point \(z^0 \in \mathcal{H}\) and \(\gamma \in \mathbb{R}_{++}\) such that if \((z^j)_{j \geq 0}\) is generated by Algorithm 1 with \(\lambda_k = 1/2\) for all \(k \geq 0\), then

\[
\begin{align*}
f(x^k_g) + g(x^k_g) - f(x^*) - g(x^*) &= \Omega \left( \frac{1}{(k + 1)^{\alpha}} \right).
\end{align*}
\]

Proof. Assume the setting of Theorem 2.7.1. Then \(f = d_V\) and \(g = \iota_U\), and by Lemma 2.7.4, we have

\[
\begin{align*}
f(x^k_g) + g(x^k_g) - f(x^*) - g(x^*) &= d_V(x_g^k) = \Omega \left( \frac{1}{(k + 1)^{\alpha}} \right).
\end{align*}
\]

Theorem 2.7.2 shows that the DRS algorithm is nearly as slow as the subgradient method. We use the word nearly because the subgradient method has complexity \(O(1/\sqrt{k + 1})\), while DRS has complexity \(o(1/\sqrt{k + 1})\). To the best of our knowledge, this is the first lower complexity result for DRS algorithm. Note that Theorem 2.7.2 implies the same lower complexity for the Forward Douglas Rachford splitting algorithm [30].

2.7.4 Optimal objective and FPR rates with Lipschitz derivative

The following examples show that the objective and FPR rates derived in Theorem 2.3.3 are essentially optimal. The setup of the following counterexample already appeared in [28, Remarque 6] but the objective function lower bounds were not shown.

Theorem 2.7.3 (Lower complexity of PPA). There exists a Hilbert space \(\mathcal{H}\), and a closed, proper, and convex function \(f\) such that for all \(\alpha > 1/2\), there exists \(z^0 \in \mathcal{H}\) such that if \((z^j)_{j \geq 0}\) is generated by PPA (Equation (2.1.8)), then

\[
\begin{align*}
\|\text{prox}_{\gamma f}(z^k) - z^k\|^2 &\geq \frac{\gamma^2}{(1 + 2\alpha)e^{2\gamma(k + \gamma)1 + 2\alpha}}; \\
f(z^{k+1}) - f(x^*) &\geq \frac{1}{4\alpha e^{2\gamma(k + 1 + \gamma)2\alpha}}.
\end{align*}
\]
Proof. Let $\mathcal{H} = \ell_2(\mathbb{R})$, and define a linear map $A : \mathcal{H} \to \mathcal{H}$:

$$A(z_1, z_2, \ldots, z_n, \ldots) = \left(z_1, \frac{z_2}{2}, \ldots, \frac{z_n}{n}, \ldots\right).$$

For all $z \in \mathcal{H}$, define $f(x) = (1/2)\langle Az, z \rangle$. Thus, we have the proximal identity for $f$ and

$$\text{prox}_{\gamma f}(z) = (I + \gamma A)^{-1}(z) = \left(\frac{j}{j + \gamma} z_j\right)_{j \geq 1} \quad \text{and} \quad (I - \text{prox}_{\gamma f})(z) = \left(\frac{\gamma}{j + \gamma} z_j\right)_{j \geq 1}.$$

Now let $z^0 = (1/(j + \gamma)^{\alpha})_{j \geq 1} \in \mathcal{H}$, and set $T = \text{prox}_{\gamma f}$. Then we get the following FPR lower bound:

$$\|z^{k+1} - z^k\|^2 = \|T^k(T - I)z^0\|^2 = \sum_{i=1}^{\infty} \left(\frac{i}{i + \gamma}\right)^{2k} \frac{\gamma^2}{(i + \gamma)^{2+2\alpha}} \geq \sum_{i=k}^{\infty} \left(\frac{i}{i + \gamma}\right)^{2k} \frac{\gamma^2}{(i + \gamma)^{2+2\alpha}} \geq \frac{\gamma^2}{(1 + 2\alpha)e^{2\gamma(k + \gamma)^{1+2\alpha}}}.$$

Furthermore, the objective lower bound holds

$$f(z^{k+1}) - f(x^*) = \frac{1}{2} \langle Az^{k+1}, z^{k+1} \rangle = \frac{1}{2} \sum_{i=1}^{\infty} \frac{1}{i} \left(\frac{i}{i + \gamma}\right)^{2(k+1)} \frac{1}{(i + \gamma)^{2\alpha}} \geq \frac{1}{2} \sum_{i=k+1}^{\infty} \left(\frac{i}{i + \gamma}\right)^{2(k+1)} \frac{1}{(i + \gamma)^{1+2\alpha}} \geq \frac{1}{4\alpha e^{2\gamma(k + 1 + \gamma)^{2\alpha}}}.$$

\[\square\]

2.8 From relaxed PRS to relaxed ADMM

It is well known that ADMM is equivalent to DRS applied to the Lagrange dual of Problem (2.1.2) [63]. Thus, if we let $d_f(w) := f^*(A^*w)$ and $d_g(w) := g^*(B^*w) - \langle w, b \rangle$, then relaxed ADMM is equivalent to relaxed PRS applied to the following problem:

$$\minimize_{w \in \mathcal{G}} d_f(w) + d_g(w). \quad (2.8.1)$$

We make two assumptions regarding $d_f$ and $d_g$: 52
Assumption 2.8.1 (Solution existence). Functions \( f, g : \mathcal{H} \rightarrow (-\infty, \infty] \) satisfy
\[
\text{zer}(\partial d_f + \partial d_g) \neq \emptyset. \tag{2.8.2}
\]

This is a restatement of Assumption 2.8.1, which we in our analysis of the primal case.

Assumption 2.8.2. The following differentiation rule holds:
\[
\partial d_f(x) = A^* \circ (\partial f^*) \circ A \quad \text{and} \quad \partial d_g(x) = B^* \circ (\partial g^*) \circ B - b.
\]

See [11, Theorem 16.37] for conditions that imply this identity, of which the weakest are \( 0 \in \text{sri}(\text{range}(A^*) - \text{dom}(f^*)) \) and \( 0 \in \text{sri}(\text{range}(B^*) - \text{dom}(g^*)) \), where sri is the strong relative interior of a convex set. This assumption may seem strong, but it is standard in the analysis of ADMM because it implies the dual proximal operator identities in (2.8.4).

Given an initial vector \( z^0 \in \mathcal{G} \), Lemma 2.4.1 shows that at each iteration relaxed PRS performs the following computations:
\[
\begin{cases}
w_{d_g}^{k+1} &= \text{prox}_{\gamma d_g}(z^k); \\
w_{d_f}^{k+1} &= \text{prox}_{\gamma d_f}(2w_{d_g}^k - z^k); \\
z^{k+1} &= z^k + 2\lambda_k (w_{d_f}^k - w_{d_g}^k). \tag{2.8.3}
\end{cases}
\]

In order to apply the relaxed PRS algorithm, we need to compute the proximal operators of the dual functions \( d_f \) and \( d_g \).

Lemma 2.8.1 (Proximity operators on the dual). Let \( w, v \in \mathcal{G} \). Then the update formulas \( w^+ = \text{prox}_{\gamma d_f}(w) \) and \( v^+ = \text{prox}_{\gamma d_g}(v) \) are equivalent to the following computations
\[
\begin{cases}
x^+ &= \text{arg min}_{x \in \mathcal{H}_1} f(x) - \langle w, Ax \rangle + \frac{\gamma}{2} \| Ax \|^2; \\
w^+ &= w - \gamma Ax^+.
\end{cases}
\]
\[
\begin{cases}
y^+ &= \text{arg min}_{y \in \mathcal{H}_2} g(y) - \langle v, By - b \rangle + \frac{\gamma}{2} \| By - b \|^2; \\
v^+ &= v - \gamma (By^+ - b). \tag{2.8.4}
\end{cases}
\]
respectively. In addition, the subgradient inclusions hold: \( A^* w^+ \in \partial f(x^+) \) and \( B^* v^+ \in \partial g(y^+) \). Finally, \( w^+ \) and \( v^+ \) are independent of the choice of \( x^+ \) and \( y^+ \), respectively, even if they are not unique solutions to the minimization subproblems.

We can use Lemma 2.8.1 to derive the relaxed form of ADMM in Algorithm 2. Note that this form of ADMM eliminates the “hidden variable” sequence \( (z^j)_{j \geq 0} \) in Equation (2.8.3). This following derivation is not new, but is included for the sake of completeness. See [63] for the original derivation.

**Proposition 2.8.1** (Relaxed ADMM). Let \( z^0 \in G \), and let \( (z^j)_{j \geq 0} \) be generated by the relaxed PRS algorithm applied to the dual formulation in Equation (2.8.1). Choose initial points \( w_{d_g}^{-1} = z^0 \), \( x^{-1} = 0 \) and \( y^{-1} = 0 \) and initial relaxation \( \lambda_{-1} = 1/2 \). Then we have the following identities starting from \( k = -1 \):

\[
y^{k+1} = \arg \min_{y \in \mathcal{H}_2} \left\{ g(y) - \langle w_{d_g}^k, Ax^k + By - b \rangle + \frac{\gamma}{2} \| Ax^k + By - b + (2\lambda_k - 1)(Ax^k + By^k - b) \|^2 \right\}
\]

\[
w_{d_g}^{k+1} = w_{d_g}^k - \gamma(Ax^k + By^{k+1} - b) - \gamma(2\lambda_k - 1)(Ax^k + By^k - b)
\]

\[
x^{k+1} = \arg \min_{x \in \mathcal{H}_1} \left\{ f(x) - \langle w_{d_g}^{k+1}, Ax + By^{k+1} - b \rangle + \frac{\gamma}{2} \| Ax + By^{k+1} - b \|^2 \right\}
\]

\[
w_{d_f}^{k+1} = w_{d_g}^{k+1} - \gamma(Ax^{k+1} + By^{k+1} - b)
\]

**Proof.** See Appendix A.1. \[ \square \]

**Remark 2.8.1.** Proposition 2.8.1 proves that \( w_{d_f}^{k+1} = w_{d_g}^{k+1} - \gamma(Ax^{k+1} + By^{k+1} - b) \). Recall that by Equation (2.8.3), \( z^{k+1} - z^k = 2\lambda_k (w_{d_f}^k - w_{d_g}^k) \). Therefore, it follows that

\[
z^{k+1} - z^k = -2\gamma\lambda_k (Ax^k + By^k - b). \tag{2.8.5}
\]

2.8.1 Dual feasibility convergence rates

We can apply the results of Section 2.5 to deduce convergence rates for the dual objective functions. Instead of restating those theorems, we just list the following bounds on the feasibility of the primal iterates.
**Theorem 2.8.1.** Suppose that \((z^j)_{j \geq 0}\) is generated by Algorithm 2, and let \((\lambda_j)_{j \geq 0} \subseteq (0, 1]\). Then the following convergence rates hold:

1. **Ergodic convergence:** The feasibility convergence rate holds:
   \[
   \|Ax^k + By^k - b\|^2 = \frac{4\|z^0 - z^*\|^2}{\gamma \lambda_k^2}.
   \] (2.8.6)

2. **Nonergodic convergence:** Suppose that \(\tau = \inf_{j \geq 0} \lambda_j (1 - \lambda_j) > 0\). Then
   \[
   \|Ax^k + By^k - b\|^2 \leq \frac{\|z^0 - z^*\|^2}{4\gamma^2 \tau (k + 1)} \quad \text{and} \quad \|Ax^k + By^k - b\|^2 = o\left(\frac{1}{k + 1}\right).
   \] (2.8.7)

**Proof.** Parts 1 and 2 are straightforward applications of Corollary 2.3.1. and the FPR identity:
\[
z^k - z^{k+1} \overset{(2.8.5)}{=} 2\gamma \lambda_k (Ax^k + By^k - b).
\] \(\square\)

### 2.8.2 Converting dual inequalities to primal inequalities

The ADMM algorithm generates 5 sequences of iterates:

\[
(z^j)_{j \geq 0}, (w^j_{d_1})_{j \geq 0}, \quad \text{and} \quad (w^j_{d_2})_{j \geq 0} \subseteq \mathcal{G} \quad \text{and} \quad (x^j)_{j \geq 0} \in \mathcal{H}_1, (y^j)_{j \geq 0} \in \mathcal{H}_2.
\]

The dual variables do not necessarily have a meaningful interpretation, so it is desirable to derive convergence rates involving the primal variables. In this section we will apply the Fenchel-Young inequality [11, Proposition 16.9] to convert the dual objective into a primal expression.

The following two propositions prove two fundamental inequalities that bound the primal objective.

**Proposition 2.8.2** (ADMM primal upper fundamental inequality). Let \(z^*\) be a fixed point of \(T_{PRS}\) and let \(w^* = \text{prox}_{\gamma d_g}(z^*)\). Then for all \(k \geq 0\), we have the bound:
\[
4\gamma \lambda_k (f(x^k) + g(y^k) - f(x^*) - g(y^*))
\]
\[
\leq \|z^k - (z^* - w^*)\|^2 - \|z^{k+1} - (z^* - w^*)\|^2 + \left(1 - \frac{1}{\lambda_k}\right) \|z^k - z^{k+1}\|^2.
\] (2.8.8)
Proof. See Appendix A.1. □

Remark 2.8.2. Note that Equation (2.8.8) is nearly identical to the upper inequality in Proposition 2.4.1, except that $z^* - w^*$ appears in the former where $x^*$ appears in the latter.

Proposition 2.8.3 (ADMM primal lower fundamental inequality). Let $z^*$ be a fixed point of $T_{PRS}$ and let $w^* = \text{prox}_{\gamma d}(z^*)$. Then for all $x \in \mathcal{H}_1$ and $y \in \mathcal{H}_2$ we have the bound:

$$f(x) + g(y) - f(x^*) - g(y^*) \geq \langle Ax + By - b, w^* \rangle.$$  \hspace{1cm} (2.8.9)

Proof. The lower bound follows from the subgradient inequalities:

$$f(x) - f(x^*) \geq \langle x - x^*, A^w w^* \rangle \quad \text{and} \quad g(y) - g(y^*) \geq \langle y - y^*, B^w w^* \rangle.$$  \hspace{1cm} (2.8.9)

We add these inequalities together and use the identity $Ax^* + By^* = b$ to get Equation (2.8.9). □

Remark 2.8.3. We use Inequality (2.8.9) in two special cases:

$$f(x^k) + g(y^k) - f(x^*) - g(y^*) \geq \frac{1}{\gamma} \langle w^k_d - w^k_d, w^* \rangle \hspace{1cm} (2.8.10)$$
$$f(x^k) + g(y^k) - f(x^*) - g(y^*) \geq \frac{1}{\gamma} \langle w^k_d - w^k_d, w^* \rangle \hspace{1cm} (2.8.11)$$

These bounds are nearly identical to the fundamental lower inequality in Proposition 2.4.2, except that $w^*$ appears in the former where $z^* - x^*$ appeared in the latter.

2.8.3 Converting dual convergence rates to primal convergence rates

In this section, we use the inequalities deduced in Section 2.8.2 to derive convergence rates for the primal objective values. The structure of the of the proofs of the following theorems are exactly the same as in the primal convergence case in Section 2.5, except that we use the upper and lower inequalities derived in the Section 2.8.2 instead of the fundamental upper and lower inequalities in Propositions 2.4.1 and 2.4.2. This amounts to replacing the term $z^* - x^*$ and $x^*$ by $w^*$ and $z^* - w^*$, respectively, in all of the inequalities from Section 2.5. Thus, we omit the proofs.
Theorem 2.8.2 (Ergodic primal convergence of ADMM). Define the ergodic primal iterates by the formulas: \( \bar{x}^k = (1/\Lambda_k) \sum_{i=0}^{k} \lambda_i x^i \) and \( \bar{y}^k = (1/\Lambda_k) \sum_{i=0}^{k} \lambda_i y^i \). Then
\[
-\frac{2\|w^*\|\|z^0 - z^*\|}{\gamma \Lambda_k} \leq f(\bar{x}^k) + g(\bar{y}^k) - f(x^*) - g(y^*) \leq \frac{\|z^0 - (z^* - w^*)\|^2}{4\gamma \Lambda_k}.
\] (2.8.12)

The ergodic rate presented here is stronger and easier to interpret than the one in [73] for the ADMM algorithm (\( \lambda_k \equiv 1/2 \)). Indeed, the rate presented in [73, Theorem 4.1] shows the following bound: for all \( k \geq 1 \) and for any bounded set \( D \subseteq \text{dom}(f) \times \text{dom}(g) \times G \), we have the following variational inequality bound
\[
\sup_{(x,y,w) \in D} \left( f(x^{k-1}) + g(y^k) - f(x) - g(y) + \langle w_{d_0}^k, Ax + By - b \rangle - \langle A\bar{x}^{k-1} + By^k - b, w \rangle \right) \leq \frac{\sup_{(x,y,w) \in D} \|(x, y, w) - (x^0, y^0, w_{d_0}^0)\|^2}{2(k+1)}.
\]
If \((x^*, y^*, w^*) \in D\), then the supremum is positive and bounds the deviation of the primal objective from the lower fundamental inequality.

Theorem 2.8.3 (Nonergodic primal convergence of ADMM). For all \( k \geq 0 \), let \( \tau_k = \lambda_k (1 - \lambda_k) \). In addition, suppose that \( \tau = \inf_{j \geq 0} \tau_j > 0 \). Then

1. In general, we have the bounds:
\[
-\frac{\|z^0 - z^*\|\|w^*\|}{2\sqrt{\tau}(k+1)} \leq f(x^k) + g(y^k) - f(x^*) - g(y^*) \leq \frac{\|z^0 - z^*\|\|z^0 - z^*\| + \|w^*\|}{2\gamma \sqrt{\tau}(k+1)}
\] (2.8.13)
and \( |f(x^k) + g(y^k) - f(x^*) - g(y^*)| = o(1/\sqrt{k+1}). \)

2. If \( G = \mathbb{R} \) and \( \lambda_k \equiv 1/2 \), then for all \( k \geq 0 \),
\[
-\frac{\|z^0 - z^*\|\|w^*\|}{\sqrt{2}(k+1)} \leq f(x^{k+1}) + g(y^{k+1}) - f(x^*) - g(x^*) \leq \frac{\|z^0 - z^*\|\|z^0 - z^*\| + \|w^*\|}{\sqrt{2}\gamma(k+1)}
\]
and \( |f(x^{k+1}) + g(x^{k+1}) - f(x^*) - g(x^*)| = o(1/(k+1)). \)
The rates presented in Theorem 2.8.3 are new and, to the best of our knowledge, they are the first nonergodic convergence rate results for ADMM primal objective error.

2.9 Examples

In this section, we apply relaxed PRS and relaxed ADMM to concrete problems and explicitly bound the associated objectives and FPR terms with the convergence rates we derived in the previous sections.

2.9.1 Feasibility problems

Suppose that \( C_f \) and \( C_g \) are closed convex subsets of \( \mathcal{H} \), with nonempty intersection. The goal of the feasibility problem is the find a point in the intersection of \( C_f \) and \( C_g \). In this section, we present one way to model this problem using convex optimization and apply the relaxed PRS algorithm to reach the minimum.

In general, we cannot expect linear convergence of relaxed PRS algorithm for the feasibility problem. We showed this in Theorem 2.6.2 by constructing an example for which the DRS iteration converges in norm but does so arbitrarily slow. A similar result holds for the alternating projection (AP) algorithm [12]. Thus, in this section we focus on the convergence rate of the FPR.

Let \( \iota_{C_f} \) and \( \iota_{C_g} \) be the indicator functions of \( C_f \) and \( C_g \). Then \( x \in C_f \cap C_g \), if, and only if, \( \iota_{C_f}(x) + \iota_{C_g}(x) = 0 \), and the sum is infinite otherwise. Thus, a point is in the intersection of \( C_f \) and \( C_g \) if, and only if, it is the minimizer of the following problem:

\[
\min_{x \in \mathcal{H}} \iota_{C_f}(x) + \iota_{C_g}(x). \tag{2.9.1}
\]

The relaxed PRS algorithm applied to this problem, with \( f = \iota_{C_f} \) and \( g = \iota_{C_g} \), has the
following form: Given $z^0 \in \mathcal{H}$, for all $k \geq 0$, let
\[
\begin{aligned}
x_g^k &= P_{C_g}(z^k); \\
x_f^k &= P_{C_f}(2x_g^k - z^k); \\
z^{k+1} &= z^k + 2\lambda_k(x_f^k - x_g^k).
\end{aligned}
\]

Because $f = \iota_{C_f}$ and $g = \iota_{C_g}$ only take on the values 0 and $\infty$, the objective value convergence rates derived earlier do not provide meaningful information, other than $x_f^k \in C_f$ and $x_g^k \in C_g$. However, from the FPR identity $x_f^k - x_g^k = 1/(2\lambda_k)(z^{k+1} - z^k)$, we find that after $k$ iterations, Corollary 2.3.1 produces the bound
\[
\max\{d_{C_g}^2(x_f^k), d_{C_f}^2(x_g^k)\} \leq \|x_f^k - x_g^k\|^2 = o\left(\frac{1}{k+1}\right)
\]
whenever $(\lambda_j)_{j \geq 0}$ is bounded away from 0 and 1. Theorem 2.6.1 showed that this rate is optimal. Furthermore, if we average the iterates over all $k$, Theorem 2.5.1 gives the improved bound
\[
\max\{d_{C_g}^2(\bar{x}_f^k), d_{C_f}^2(\bar{x}_g^k)\} \leq \|\bar{x}_f^k - \bar{x}_g^k\|^2 = O\left(\frac{1}{\Lambda_k^2}\right),
\]
which is optimal by Proposition 2.7.1. Note that the averaged iterates satisfy $\bar{x}_f^k = (1/\Lambda_k) \sum_{i=0}^{k} \lambda_i x_f^i \in C_f$ and $\bar{x}_g^k = (1/\Lambda_k) \sum_{i=0}^{k} \lambda_i x_g^i \in C_g$, because $C_f$ and $C_g$ are convex. Thus, we can state the following proposition:

**Proposition 2.9.1.** After $k$ iterations the relaxed PRS algorithm produces a point in each set with distance of order $O(1/\Lambda_k)$ from each other.

### 2.9.2 Parallelized model fitting and classification

The following general scenario appears in [26, Chapter 8]. Consider the following general convex model fitting problem: Let $M : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a feature matrix, let $b \in \mathbb{R}^m$ be the output vector, let $l : \mathbb{R}^m \rightarrow (-\infty, \infty]$ be a loss function and let $r : \mathbb{R}^n \rightarrow (-\infty, \infty]$ be a regularization function. The model fitting problem is formulated as the following
minimization:

\[
\text{minimize } l(Mx - b) + r(x).
\]  \hspace{1cm} (2.9.5)

The function \( l \) is used to enforce the constraint \( Mx = b + \nu \) up to some noise \( \nu \) in the measurement, while \( r \) enforces the \textit{regularity} of \( x \) by incorporating \textit{prior knowledge} of the form of the solution. The function \( r \) can also be used to enforce the uniqueness of the solution of \( Mx = b \) in ill-posed problems.

We can solve Equation (2.9.5) by a direct application of relaxed PRS and obtain \( O(1/\Lambda_k) \) ergodic convergence and \( o \left( 1/\sqrt{k + 1} \right) \) nonergodic convergence rates. Note that these rates do not require differentiability of \( f \) or \( g \). In contrast, the FBS algorithm requires differentiability of one of the objective functions and a knowledge of the Lipschitz constant of its gradient. The advantage of FBS is the \( o(1/(k+1)) \) convergence rate shown in Theorem 2.3.3. However, we do not necessarily assume that \( l \) is differentiable, so we may need to compute \( \text{prox}_{\gamma l(M(\cdot) - b)} \), which can be significantly more difficult than computing \( \text{prox}_{\gamma l} \). Thus, in this section we separate \( M \) from \( l \) by rephrasing Equation (2.9.5) in the form of Problem (2.1.2).

In this section, we present several different ways to split Equation (2.9.5). Each splitting gives rise to a different algorithm and can be applied to general convex \( l \) and \( r \). Our results predict convergence rates that hold for primal objectives, dual objectives, and the primal feasibility. Note that in parallelized model fitting, it is not always desirable to take the time average of all of the iterates. Indeed, when \( r \) enforces sparsity, averaging the current \( r \)-iterate with old iterates, all of which are sparse, can produce a non-sparse iterate. This will slow down vector additions and prolong convergence.
2.9.2.1 Auxiliary variable

We can split Equation (2.9.5) by defining an auxiliary variable for $My$:

\[
\begin{align*}
\text{minimize} \quad & f(x) + g(y) \\
\text{subject to} \quad & My - x = b.
\end{align*}
\]

The constraint in Equation (2.9.6) reduces to $Ax + By = b$ where $B = M$ and $A = -I_{R^n}$. If we set $f = l$ and $g = r$ and apply ADMM, the analysis of Section 2.8.3 shows that

\[
\begin{align*}
|l(x^k) + r(y^k) - l(My^k - b) - r(y^*)| &= o\left(\frac{1}{\sqrt{k+1}}\right); \\
\|My^k - b - x^k\|^2 &= o\left(\frac{1}{k+1}\right).
\end{align*}
\]

In particular, if $l$ is Lipschitz, then $|l(x^k) - l(My^k - b)| = o\left(\frac{1}{\sqrt{k+1}}\right)$. Thus, we have

\[
|l(My^k - b) + r(y^k) - l(My^* - b) - r(y^*)| = o\left(\frac{1}{\sqrt{k+1}}\right).
\]

A similar analysis shows that

\[
\begin{align*}
|l(My^k - b) + r(y^k) - l(My^* - b) - r(y^*)| &= O\left(\frac{1}{\Lambda^k}\right); \\
\|My^k - b - x^k\|^2 &= O\left(\frac{1}{\Lambda^2_k}\right).
\end{align*}
\]

In the following two splittings, we leave the derivation of convergence rates to the reader.

2.9.2.2 Splitting across examples

We assume that $l$ is block separable: we have $l(Mx - b) = \sum_{i=1}^{R} l_{i}(M_{i}x - b_{i})$ where

\[
M = \begin{bmatrix}
M_{1} \\
\vdots \\
M_{R}
\end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix}
b_{1} \\
\vdots \\
b_{R}
\end{bmatrix}.
\]
Each $M_i \in \mathbb{R}^{m_i \times n}$ is a submatrix of $M$, each $b_i \in \mathbb{R}^{m_i}$ is a subvector of $b$, and $\sum_{i=1}^{R} m_i = m$. Therefore, an equivalent form of Equation (2.9.5) is given by

$$\begin{align*}
\text{minimize} & \quad x, \cdots, x_R, y \in \mathbb{R}^n \\
\text{subject to} & \quad x_r - y = 0, \quad r = 1, \cdots, R.
\end{align*}$$

We say that Equation (2.9.7) is split across examples. Thus, to apply ADMM to this problem, we simply stack the vectors $x_i, i = 1, \cdots, R$ into a vector $x = (x_1, \cdots, x_R)^T \in \mathbb{R}^{nR}$. Then the constraints in Equation (2.9.7) reduce to $Ax + By = 0$ where $A = I_{\mathbb{R}^{nR}}$ and $By = (-y, \cdots, -y)^T$.

### 2.9.2.3 Splitting across features

We can also split Equation (2.9.5) across features, whenever $r$ is block separable in $x$, in the sense that there exists $C > 0$, such that $r = \sum_{i=1}^{C} r_i(x_i)$, and $x_i \in \mathbb{R}^{n_i}$ where $\sum_{i=1}^{C} n_i = n$. This splitting corresponds to partitioning the columns of $M$, i.e. $M = \begin{bmatrix} M_1, \cdots, M_C \end{bmatrix}$, and $M_i \in \mathbb{R}^{m \times n_i}$, for all $i = 1, \cdots, C$. Note that for all $y \in \mathbb{R}^n$, $My = \sum_{i=1}^{C} M_i y_i$. With this notation, we can derive an equivalent form of Equation (2.9.5) given by

$$\begin{align*}
\text{minimize} & \quad l \left( \sum_{i=1}^{C} x_i - b \right) + \sum_{i=1}^{C} r_i(y_i) \\
\text{subject to} & \quad x_i - M_i y_i = 0, \quad i = 1, \cdots, C.
\end{align*}$$

The constraint in Equation (2.9.8) reduces to $Ax + By = 0$ where $A = I_{\mathbb{R}^{nC}}$ and $By = -(M_1 y_1, \cdots, M_C y_C)^T \in \mathbb{R}^{nC}$.

### 2.9.3 Distributed ADMM

In this section our goal is to use Algorithm 2 to

$$\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{m} f_i(x)
\end{align*}$$

62
by using the splitting in [107]. Note that we could minimize this function by reformulating it in the product space $\mathcal{H}^m$ as follows:

$$\minimize_{x \in \mathcal{H}^m} \sum_{i=1}^{m} f_i(x_i) + \iota_D(x),$$

where $D = \{(x, \cdots, x) \in \mathcal{H}^m \mid x \in \mathcal{H}\}$ is the diagonal set. Applying relaxed PRS to this problem results in a parallel algorithm where each function performs a local minimization step and then communicates its local variable to a central processor. In this section, we assign each function a local variable but we never communicate it to a central processor. Instead, each function only communicates with neighbors.

Formally, we assume there is a simple, connected and undirected graph $G = (V, E)$ on $|V| = m$ vertices with edges, $E$, that describe a neighbor relation among the different functions. We introduce a new variable $x_i \in \mathcal{H}$ for each function $f_i$, and, hence, we set $\mathcal{H}_1 = \mathcal{H}^m$, (see Section 2.8). We can encode the constraint that each node communicates with neighbors by introducing an auxiliary variable for each edge in the graph:

$$\minimize_{x \in \mathcal{H}^m, y \in \mathcal{H}^{|E|}} \sum_{i=1}^{m} f_i(x_i)$$

subject to $x_i = y_{ij}, x_j = y_{ij}$, for all $(i, j) \in E$.  

The linear constraints in Equation (2.9.11) can be written in the form of $A x + B y = 0$ for proper matrices $A$ and $B$. Thus, we reformulate Equation (2.9.11) as

$$\minimize_{x \in \mathcal{H}^m, y \in \mathcal{H}^{|E|}} \sum_{i=1}^{m} f_i(x_i) + g(y)$$

subject to $A x + B y = 0$,  

where $g : \mathcal{H}^{|E|} \to \mathbb{R}$ is the zero map.

Because we only care about finding the value of the variable $x \in \mathcal{H}^m$, the following simplification can be made to the sequences generated by ADMM applied to Equation (2.9.12) with $\lambda_k = 1/2$ for all $k \geq 1$ [110]: Let $\mathcal{N}_i$ denote the set of neighbors of $i \in V$ and set
\( x_i^0 = \alpha_i^0 = 0 \) for all \( i \in V \). Then for all \( k \geq 0 \),

\[
\begin{align*}
\begin{cases}
x_i^{k+1} = \arg \min_{x_i \in \mathcal{H}} f_i(x) + \frac{\gamma |N_i|}{2} \|x_i - x_i^k - \frac{1}{|N_i|} \sum_{j \in N_i} x_j^k + \frac{1}{\gamma |N_i|} \alpha_i \|^2 + \frac{\gamma |N_i|}{2} \|x_i\|^2
\end{cases}
\end{align*}
\]

\[\alpha_i^{k+1} = \alpha_i^k + \gamma \left( |N_i| x_i^{k+1} - \sum_{j \in N_i} x_j^{k+1} \right).\]

Equation (2.9.13) is truly distributed because each node \( i \in V \) only requires information from its local neighbors at each iteration.

In [110], linear convergence is shown for this algorithm provided that \( f_i \) are strongly convex and \( \nabla f_i \) are Lipschitz. For general convex functions, we can deduce the nonergodic rates from Theorem 2.8.3

\[
\left| \sum_{i=1}^{m} f_i(x_i^k) - f(x^*) \right| = o \left( \frac{1}{\sqrt{k+1}} \right);
\]

\[
\sum_{i \in V j \in N_i} \|x_i^k - z_{ij}^k\|^2 + \sum_{i \in V i \in N_j} \|x_j^k - z_{ij}^k\|^2 = o \left( \frac{1}{k+1} \right),
\]

and the ergodic rates from Theorem 2.8.2

\[
\left| \sum_{i=1}^{m} f_i(\bar{x}_i^k) - f(x^*) \right| = O \left( \frac{1}{k+1} \right);
\]

\[
\sum_{i \in V j \in N_i} \|\bar{x}_i^k - z_{ij}^k\|^2 + \sum_{i \in V i \in N_j} \|\bar{x}_j^k - z_{ij}^k\|^2 = O \left( \frac{1}{(k+1)^2} \right).
\]

These convergence rates are new and complement the linear convergence results in [110]. In addition, they complement the similar ergodic rate derived in [114] for a different distributed splitting.

### 2.10 Conclusion

In this chapter, we provided a comprehensive convergence rate analysis of the FPR and objective error of several splitting algorithms under general convexity assumptions. We showed that the convergence rates are essentially optimal in all cases. All results follow
from some combination of a lemma that deduces convergence rates of summable monotonic sequences (Lemma 2.2.1), a simple diagram (Figure 2.1), and fundamental inequalities (Propositions 2.4.1 and 2.4.2) that relate the FPR to the objective error of the relaxed PRS algorithm. The most important open question is whether and how the rates we derived will improve when we enforce stronger assumptions, such as Lipschitz differentiability and/or strong convexity, on $f$ and $g$. This will be the subject of future work.
CHAPTER 3

Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions

3.1 Introduction

The Douglas-Rachford splitting (DRS), Peaceman-Rachford splitting (PRS), and alternating direction method of multipliers (ADMM) algorithms are abstract splitting schemes that solve monotone inclusion and convex optimization problems [85, 65, 64]. The DRS and PRS algorithms solve monotone inclusion problems in which the operator is the sum of two (possibly) simpler operators by accessing each operator individually through its resolvent. The ADMM algorithm solves convex optimization problems in which the objective is the sum of two (possibly) simpler functions with variables linked through a linear constraint via an alternating minimization strategy. The variable splitting that occurs in each of these algorithms can give rise to parallel and even distributed implementations of minimization algorithms [26, 110, 114], which are particularly suitable for large-scale applications. Since the 1950s, these methods were largely applied to solving partial differential equations (PDEs) and feasibility problems, and only recently has their power been utilized in (PDE and non-PDE related) image processing, statistical and machine learning, compressive sensing, matrix completion, finance, and control [68, 26].

In this chapter, we consider two prototype optimization problems: the unconstrained
problem

$$\minimize_{x \in \mathcal{H}} f(x) + g(x)$$ \hspace{1cm} (3.1.1)

where $\mathcal{H}$ is a Hilbert space, and the linearly constrained variant

$$\minimize_{x \in \mathcal{H}_1, \ y \in \mathcal{H}_2} f(x) + g(y)$$

subject to $Ax + By = b$ \hspace{1cm} (3.1.2)

where $\mathcal{H}_1, \mathcal{H}_2$, and $\mathcal{G}$ are Hilbert spaces, the vector $b$ is an element of $\mathcal{G}$, and $A: \mathcal{H}_1 \to \mathcal{G}$ and $B: \mathcal{H}_2 \to \mathcal{G}$ are linear operators. Our working assumption throughout the chapter is that the subproblems involving $f$ and $g$ separately are much simpler to solve than the joint minimization problem.

Although Problem (3.1.1) is a special case of Problem (3.1.2), we consider them separately because the algorithms used to solve each problem and the target application area are usually different. Problem (3.1.1) models a variety of tasks in signal recovery where one function corresponds to a data fitting term and the other enforces prior knowledge, such as sparsity, low rank, or smoothness [46]. In this chapter, we apply relaxed PRS (Algorithm 4) to solve Problem (3.1.1). On the other hand, Problem (3.1.2) models tasks in machine learning, image processing and distributed optimization. The linear constraint can be used to enforce data fitting, but it can also be used to split variables in a way that gives rise to parallel or distributed optimization algorithms [18, 26]. In this chapter, we apply relaxed ADMM (Algorithm 5) to Problem (3.1.2).

### 3.1.1 Goals, challenges, and approaches

This work seeks to improve the theoretical understanding of DRS, PRS, and ADMM, as well as their averaged versions. When applied to convex optimization problems, they are known to converge under rather general conditions [11, Corollary 27.4]. The objective error convergence rates and worst case lower complexity of these algorithms were analyzed under general convexity assumptions in Chapter 2. This work seeks to complement the
results of Chapter 2 by deriving stronger rates under correspondingly stronger conditions on Problems 3.1.1 and 3.1.2. One of the main consequences of this work is that the relaxed PRS and ADMM algorithms automatically adapt to the regularity of the problem at hand and achieve convergence rates that improve upon the (tight) worst-case rates shown in Chapter 2 for the nonsmooth case. Thus, our results offer an explanation of the great performance of relaxed PRS and ADMM observed in practice, and together with Chapter 2 we now have a comprehensive convergence rate analysis of the relaxed PRS and ADMM algorithms.

In this chapter, we derive the convergence rates of the objective error and fixed-point residual (FPR) of the relaxed PRS applied to Problem (3.1.1). In addition, we derive the convergence rates of constraint violations, the primal objective error, and the dual objective error for relaxed ADMM applied to Problem (3.1.2). The derived rates are useful for determining how many iterations are needed to reach a certain accuracy, to decide when to stop an algorithm, and to compare relaxed PRS and ADMM to other algorithms in terms of their worst-case complexities.

We now describe our contributions and techniques:

i We show that if $f$ or $g$ is strongly convex, then a natural sequence of points converges strongly, the best iterate converges with rate $o(1/(k + 1))$, and the ergodic iterate converges with rate $O(1/(k + 1))$ (Theorem 3.2.1). The proofs follow by showing that a certain sequence of squared norms is summable. This result is in stark contrast to Theorem 2.6.2 of Chapter 2, which shows that DRS can converge arbitrarily slowly when strong convexity does not hold.

ii We show that if $f$ or $g$ has a Lipschitz derivative, the best objective error after $k$ iterations has order $o(1/(k + 1))$ for any choice of input parameters (Theorem 3.3.2). This rate is in stark contrast to the nonsmooth case where the convergence rate $o(1/\sqrt{k + 1})$ is tight (see Theorem 2.7.2 of Chapter 2). The result follows by showing the objective error is summable.
iii We show that if the function $g$ has a Lipschitz derivative, the implicit stepsize parameter $\gamma$ is chosen small enough, and the relaxation parameters satisfy $\lambda_k \equiv 1/2$, then the objective error has order $o(1/(k + 1))$ (Theorem 3.3.3). Furthermore, under the same assumptions, we show that the FPR has order $o(1/(k + 1)^2)$ (Theorem 3.3.4). We conclude that the DRS algorithm is always at least as fast as the forward-backward splitting (FBS) algorithm. (See [14] for the big-$O$ FBS rate and Theorem 2.3.3 of Chapter 2 for the little-$o$ FBS rate.) The results follow by showing that a sequence that dominates the objective error is monotonic and summable. The derived rates are shown to be optimal by an example in Theorem 2.7.3 in Chapter 2.

iv We prove that the relaxed PRS algorithm converges linearly whenever at least one of the objectives is strongly convex and at least one has a Lipschitz derivative (Section 3.4).

v We show that a collection of projection splitting algorithms, which contain the method of alternating projections (MAP) as a special case, converge linearly when the underlying sets have a nice intersection (Section 3.5). Our proof shows that these algorithms are a special case of the relaxed PRS algorithm. We recover several classical results and derive several new convergence rates for algorithms that have not appeared in the literature.

vi We give the convergence rates of primal objective error and feasibility measures in ADMM by applying the Fenchel-Young inequality and extending the above results (Section 3.6).

Much of our analysis is built on Chapter 2, where several splitting schemes are analyzed under general convexity assumptions. The following sections contain a brief review of the main results that we utilize from Chapter 2.
3.1.2 Notation

In what follows, \( \mathcal{H}, \mathcal{H}_1, \mathcal{H}_2, \mathcal{G} \) denote (possibly infinite dimensional) Hilbert spaces. In fixed-point iterations, \((\lambda_j)_{j \geq 0} \subseteq \mathbb{R}_+\) will denote a sequence of relaxation parameters, and

\[
\Lambda_k := \sum_{i=0}^{k} \lambda_i
\]

(3.1.3)

is its \( k \)th partial sum. To ease notational memory, the reader may assume that \( \lambda_k \equiv (1/2) \) and \( \Lambda_k = (k + 1)/2 \) in the DRS algorithm, or that \( \lambda_k \equiv 1 \) and \( \Lambda_k = (k + 1) \) in the PRS algorithm. Given the sequence \((x^j)_{j \geq 0} \subseteq \mathcal{H}\), we let \( \bar{x}^k = (1/\Lambda_k) \sum_{i=0}^{k} \lambda_i x^i \) denote its \( k \)th average with respect to the sequence \((\lambda_j)_{j \geq 0}\).

We call a convergence result *ergodic* if it applies to the sequence \((x^j)_{j \geq 0}\), and *nonergodic* if it applies to the sequence \((x^j)_{j \geq 0}\).

For any subset \( C \subseteq \mathcal{H} \), we define the distance function:

\[
d_C(x) := \inf_{y \in C} \|x - y\|
\]

(3.1.4)

Given a closed, proper, and convex function \( f : \mathcal{H} \to (-\infty, \infty] \), the set \( \partial f(x) \) denotes its subdifferential at \( x \) and

\[
\tilde{\nabla} f(x) \in \partial f(x)
\]

denotes a subgradient. (This notation was used in [16, Eq. (1.10)].)

The convex conjugate of a closed, proper, and convex function \( f \) is

\[
f^*(y) := \sup_{x \in \mathcal{H}} \langle y, x \rangle - f(x).
\]

Let \( I_\mathcal{H} : \mathcal{H} \to \mathcal{H} \) denote the identity map. For any point \( x \in \mathcal{H} \) and scalar \( \gamma \in \mathbb{R}_+ \), we let

\[
\text{prox}_{\gamma f}(x) := \arg \min_{y \in \mathcal{H}} f(y) + \frac{1}{2\gamma} \|y - x\|^2 \quad \text{and} \quad \text{refl}_{\gamma f} := 2\text{prox}_{\gamma f} - I_\mathcal{H},
\]

70
which are known as the proximal and reflection operators. In addition, we define the PRS operator:

\[ T_{PRS} := \text{refl}_{\gamma f} \circ \text{refl}_{\gamma g}. \]

Let \( \lambda > 0 \). For every nonexpansive map \( T : \mathcal{H} \to \mathcal{H} \) we use the notation:

\[ T_{\lambda} := (1 - \lambda)I_{\mathcal{H}} + \lambda T. \]

We call the following identity the cosine rule:

\[ \|y - z\|^2 + 2\langle y - x, z - x \rangle = \|y - x\|^2 + \|z - x\|^2, \quad \forall x, y, z \in \mathcal{H}. \] (3.1.5)

### 3.1.3 Assumptions

We list the assumptions used throughout this chapter as follows.

**Assumption 3.1.1** (Problem assumptions). Every function we consider is closed, proper, and convex.

Unless otherwise stated, a function is not necessarily differentiable.

**Assumption 3.1.2** (Solution existence). Functions \( f, g : \mathcal{H} \to (-\infty, \infty] \) satisfy

\[ \text{zer}(\partial f + \partial g) \neq \emptyset. \] (3.1.6)

Note that this assumption is slightly stronger than the existence of a minimizer because \( \text{zer}(\partial f + \partial g) \neq \text{zer}(\partial(f + g)) \), in general [11, Remark 16.7]. Nevertheless, this assumption is standard.

**Assumption 3.1.3** (Differentiability). Every differentiable function we consider is Fréchet differentiable [11, Definition 2.45].
3.1.4 The Algorithms

The results of this chapter apply to several operator-splitting algorithms that are all based on the atomic evaluation of the proximal operator. By default, all algorithms start from an arbitrary \( z^0 \in \mathcal{H} \). The Douglas-Rachford splitting (DRS) algorithm applied to minimizing \( f + g \) is as follows:

\[
\begin{align*}
    x^k_g &= \text{prox}_{\gamma g}(z^k); \\
    x^k_f &= \text{prox}_{\gamma f}(2x^k_g - z^k); \quad k = 0, 1, \ldots, \\
    z^{k+1} &= z^k + (x^k_f - x^k_g);
\end{align*}
\]

which has the equivalent operator-theoretic and subgradient form (Lemma 3.1.2):

\[
z^{k+1} = \frac{1}{2}(I_{\mathcal{H}} + T_{\text{PRS}})(z^k) = z^k - \gamma(\tilde{\nabla}f(x^k_f) + \tilde{\nabla}g(x^k_g)), \quad k = 0, 1, \ldots,
\]

where \( \tilde{\nabla}f(x^k_f) \in \partial f(x^k_f) \) and \( \tilde{\nabla}g(x^k_g) \in \partial g(x^k_g) \). (See Part 1 of Proposition 3.1.1 for how the notation \( \tilde{\nabla} \) relates to \( \text{prox} \).) In the above algorithm, we can replace the \((1/2)\)-average of \( I_{\mathcal{H}} \) and \( T_{\text{PRS}} \) with any other weight; this results the relaxed PRS algorithm:

**Algorithm 4: Relaxed Peaceman-Rachford Splitting (relaxed PRS)**

**input**: \( z^0 \in \mathcal{H}, \gamma > 0, (\lambda_j)_{j \geq 0} \subset (0, 1] \)

**for** \( k = 0, 1, \ldots \) **do**

\[
z^{k+1} = (1 - \lambda_k)z^k + \lambda_k \text{refl}_{\gamma f} \circ \text{refl}_{\gamma g}(z^k);
\]

The special cases \( \lambda_k \equiv 1/2 \) and \( \lambda_k \equiv 1 \) are called the DRS and PRS algorithms, respectively.

The relaxed PRS algorithm can be applied to problem (3.1.2). To this end we define the Lagrangian:

\[
    \mathcal{L}_\gamma(x, y; w) := f(x) + g(y) - \langle w, Ax + By - b \rangle + \frac{\gamma}{2}||Ax + By - b||^2.
\]

Section 3.6 presents Algorithm 4 applied to the Lagrange dual of (3.1.2), which reduces
to the following algorithm:

**Algorithm 5**: Relaxed alternating direction method of multipliers (relaxed ADMM)

input : \( w^{-1} \in H, x^{-1} = 0, y^{-1} = 0, \lambda_{-1} = \frac{1}{2}, \gamma > 0, (\lambda_j)_{j \geq 0} \subseteq (0, 1] \)

for \( k = -1, 0, \ldots \) do

\[
\begin{align*}
   y^{k+1} &= \text{arg min}_y \mathcal{L}_\gamma(x^k, y; w^k) + \gamma(2\lambda_k - 1)(By, (Ax^k + By^k - b)); \\
   w^{k+1} &= w^k - \gamma(Ax^k + By^{k+1} - b) - \gamma(2\lambda_k - 1)(Ax^k + By^k - b); \\
   x^{k+1} &= \text{arg min}_x \mathcal{L}_\gamma(x, y^{k+1}; w^{k+1});
\end{align*}
\]

If \( \lambda_k \equiv 1/2 \), Algorithm 5 recovers the standard ADMM.

### 3.1.5 Practical implications: a comparison with FBS

Suppose that the function \( g \) in Problem 3.1.1 is differentiable and \( \nabla g \) is \((1/\beta)-\text{Lipschitz}\). Under this smoothness assumption, we can apply FBS algorithm to Problem 3.1.1: given \( z^0 \in H \), for all \( k \geq 0 \), define

\[
z^{k+1} = \text{prox}_{\gamma f}(z^k - \gamma \nabla g(z^k)).
\]

Note that in order to ensure convergence of the FBS algorithm, the implicit stepsize parameter \( \gamma \) must be strictly less than \( 2\beta \).

Now because the gradient operator is often simpler to evaluate than the proximal operator, it may be preferable to use FBS instead of relaxed PRS whenever one of the objectives is differentiable. From our results, we can give two reasons why it may be preferable to use relaxed PRS over FBS:

1. If the Lipschitz constant of the gradient is known, our analysis indicates how to properly choose stepsizes of relaxed PRS so that both algorithms converge with the same rate (Theorem 3.3.3). In practice, relaxed PRS is often observed to converge faster than FBS, so our results at least indicate that we can do no worse by using relaxed PRS.

2. If the Lipschitz constant of the gradient is not known, a line search procedure...
can be used to guarantee convergence of FBS. If this procedure is more expensive
than evaluating the proximal operator, then relaxed PRS should be used. Indeed,
Theorem 3.3.2 shows that the “best iterate” of relaxed PRS will converge with rate
\( o(1/(k+1)) \) regardless of the chosen stepsize, whereas FBS may fail to converge.

Thus, one of our main contributions is the “demystification” of parameter choices, and
a partial explanation of the perceived practical advantage of relaxed PRS over FBS.

3.1.6 Basic properties of proximal operators

The following properties are included in textbooks such as [11].

Proposition 3.1.1. Let \( f, g : \mathcal{H} \to (-\infty, \infty) \) be closed, proper, and convex functions,
and let \( T : \mathcal{H} \to \mathcal{H} \) be nonexpansive. The following are true:

1. **Optimality conditions of prox:** Let \( x \in \mathcal{H} \). Then \( x^+ = \text{prox}_{\gamma f}(x) \) if, and only
   if,
   \[
   \tilde{\nabla} f(x^+) := \frac{1}{\gamma}(x - x^+) \in \partial f(x^+).
   \]

2. **Proximal operators are \( 1/2 \)-averaged:** The operator \( \text{prox}_{\gamma f} : \mathcal{H} \to \mathcal{H} \) satisfies
   the following contraction property:
   \[
   \|\text{prox}_{\gamma f}(x) - \text{prox}_{\gamma f}(y)\| \leq \|x - y\| - \|(x - \text{prox}_{\gamma f}(x)) - (y - \text{prox}_{\gamma f}(y))\|.
   \]
   (3.1.7)

3. **Nonexpansiveness of the PRS operator:** The operator \( \text{refl}_{\gamma f} : \mathcal{H} \to \mathcal{H} \) is
   nonexpansive. Therefore, the composition is nonexpansive:
   \[
   T_{\text{PRS}} := \text{refl}_{\gamma f} \circ \text{refl}_{\gamma g}.
   \]
   (3.1.8)

3.1.7 Convergence rates of summable sequences

The following lemma will be key to deducing Convergence rates in Sections 3.2 and 3.3.
For the readers benefit, we recall the statement from Chapter 2.
Lemma 3.1.1 (Summable sequence convergence rates). Suppose that the nonnegative scalar sequences \((\lambda_j)_{j \geq 0}\) and \((a_j)_{j \geq 0}\) satisfy \(\sum_{i=0}^{\infty} \lambda_i a_i < \infty\). Let \(\Lambda_k := \sum_{i=0}^{k} \lambda_i\) for \(k \geq 0\).

1. **Monotonicity:** If \((a_j)_{j \geq 0}\) is monotonically nonincreasing, then
   
   \[
   a_k \leq \frac{1}{\Lambda_k} \left( \sum_{i=0}^{\infty} \lambda_i a_i \right) \quad \text{and} \quad a_k = o\left( \frac{1}{\Lambda_k - \Lambda_{\lceil k/2 \rceil}} \right). \tag{3.1.9}
   
   In particular,

   (a) if \((\lambda_j)_{j \geq 0}\) is bounded away from 0, then \(a_k = o(1/(k+1))\);

   (b) if \(\lambda_k = (k+1)^p\) for some \(p \geq 0\) and all \(k \geq 1\), then \(a_k = o(1/(k+1)^{p+1})\);

   (c) as a special case, if \(\lambda_k = (k+1)\) for all \(k \geq 0\), then \(a_k = o(1/(k+1)^2)\).

2. **Monotonicity up to errors:** Let \((e_j)_{j \geq 0}\) be a sequence of scalars. Suppose that \(a_{k+1} \leq a_k + e_k\) for all \(k\) (where \(e_k\) represents an error) and that \(\sum_{i=0}^{\infty} \Lambda_i e_i < \infty\). Then
   
   \[
   a_k \leq \frac{1}{\Lambda_k} \left( \sum_{i=0}^{\infty} \lambda_i a_i + \sum_{i=0}^{\infty} \Lambda_i e_i \right) \quad \text{and} \quad a_k = o\left( \frac{1}{\Lambda_k - \Lambda_{\lceil k/2 \rceil}} \right). \tag{3.1.10}
   
   The rates of \(a_k\) in Parts 1a, 1b, and 1c continue to hold as long as \(\sum_{i=0}^{\infty} \Lambda_i e_i < \infty\) holds. In particular, they hold if \(e_k = O(1/(k+1)^q)\) for some \(q > 2\), \(q > p + 2\), and \(q > 3\), respectively.

3. **Faster rates:** Suppose \((b_j)_{j \geq 0}\) and \((e_j)_{j \geq 0}\) are nonnegative scalar sequences, that \(\sum_{i=0}^{\infty} b_j < \infty\) and \(\sum_{i=0}^{\infty} (i+1)e_i < \infty\), and that for all \(k \geq 0\) we have \(\lambda_k a_k \leq b_k - b_{k+1} + e_k\). Then the following sum is finite:
   
   \[
   \sum_{i=0}^{\infty} (i+1)\lambda_i a_i \leq \sum_{i=0}^{\infty} b_i + \sum_{i=0}^{\infty} (i+1)e_i < \infty. \tag{3.1.11}
   
   In particular,

   (a) if \((\lambda_j)_{j \geq 0}\) is bounded away from 0, then \(a_k = o(1/(k+1)^2)\);

   (b) if \(\lambda_k = (k+1)^p\) for some \(p \geq 0\) and all \(k \geq 1\), then \(a_k = o(1/(k+1)^{p+2})\).
4. **No monotonicity:** For all $k \geq 0$, define the sequence of indices

$$k_{\text{best}} := \arg \min_i \{ a_i | i = 0, \ldots, k \}.$$ 

Then $(a_{j_{\text{best}}})_{j \geq 0}$ is monotonically nonincreasing and the above bounds continue to hold when $a_k$ is replaced with $a_{k_{\text{best}}}$.

### 3.1.8 Convergence of the fixed-point residual (FPR)

In this section, we note a few key results about the operator $T_{\text{PRS}}$ that will be useful in the later sections. For the readers convenience, we recall the following result from Theorem 2.3.1 of Chapter 2

**Proposition 3.1.2.** Let $z^* \in \mathcal{H}$ be a fixed point of $T_{\text{PRS}}$, and let $(z^i)_{i \geq 0}$ be generated by the relaxed PRS algorithm:

$$z^{k+1} = (T_{\text{PRS}})_{\lambda_k} z^k.$$ 

Then the following are true:

1. $(\|z^i - z^*\|)_{i \geq 0}$ is monotonically nonincreasing;

2. $(\|T_{\text{PRS}} z^i - z^i\|)_{i \geq 0}$ is monotonically nonincreasing, and thus so is $(1/\lambda_j)\|z^{i+1} - z^j\|)_{j \geq 0}$;

3. If $\lambda_k \equiv \lambda$, then $(\|T_{\text{PRS}})_{\lambda z^i} - z^*\|)_{i \geq 0}$ is monotonically nonincreasing;

4. The Fejér-type inequality holds: for all $\lambda \in [0, 1]$

$$\|(T_{\text{PRS}})_{\lambda} z^k - z^*\|^2 \leq \|z^k - z^*\|^2 - \frac{1 - \lambda}{\lambda} \|(T_{\text{PRS}})_{\lambda} z^k - z^k\|^2. \quad (3.1.12)$$

5. For all $k \geq 0$, let $\tau_k = \lambda_k (1 - \lambda_k)$. Then

$$\sum_{i=0}^{\infty} \tau_i \|T_{\text{PRS}} z^i - z^i\| \leq \|z^0 - z^*\|^2.$$
6. If $\tau := \inf_{j \geq 0} \lambda_j (1 - \lambda_j) > 0$, then the following convergence rates hold:

$$\|T_{\text{PRS}} z^k - z^k\|^2 \leq \frac{\|z^0 - z^*\|^2}{\tau (k + 1)} \quad \text{and} \quad \|T_{\text{PRS}} z^k - z^k\|^2 = o \left( \frac{1}{\tau (k + 1)} \right).$$

(3.1.13)

**Remark 3.1.1.** We call the quantity

$$\|T_{\text{PRS}} z^k - z^k\|^2$$

the fixed-point residual (FPR) of the relaxed PRS algorithm. Throughout this chapter, we slightly abuse terminology and call the successive iterate difference $\|z^{k+1} - z^k\|^2 = \lambda_k^2 \|T_{\text{PRS}} z^k - z^k\|^2$ FPR as well.

### 3.1.9 Subgradients

For the readers convenience, we recall Figure 3.1 which also appeared as Figure 2.1 in Chapter 2. This diagram is key to deducing all of the algebraic relations necessary for relating the objective error to the FPR of the relaxed PRS iteration.

![Figure 3.1: A single relaxed PRS iteration starting from $z$.](image)

**Figure 3.1:** A single relaxed PRS iteration starting from $z$.  

Lemma 3.1.2 summarizes the identities depicted in Figure 3.1.
Lemma 3.1.2. Let $z \in \mathcal{H}$. Define auxiliary points $x_g := \text{prox}_{\gamma g}(z)$ and $x_f := \text{prox}_{\gamma f}(\text{refl}_{g}(z))$. Then the identities hold:

$$x_g = z - \gamma \nabla g(x_g) \quad \text{and} \quad x_f = x_g - \gamma \nabla g(x_g) - \gamma \nabla f(x_f). \quad (3.1.14)$$

In addition, each relaxed PRS step $z^+ = (T_{PRS})_\lambda(z)$ has the following representation:

$$z^+ - z = 2\lambda (x_f - x_g) = -2\lambda \gamma (\nabla g(x_g) + \nabla f(x_f)). \quad (3.1.15)$$

The following optimality conditions are well known. They will be needed in Section 3.5 because we vary the implicit stepsize parameter $\gamma$. See Lemma 2.4.1 of Chapter 2 or [11] for a proof.

Lemma 3.1.3 (Optimality conditions of $T_{PRS}$). The set of zeros of $\partial f + \partial g$ is precisely

$$\text{zer}(\partial f + \partial g) = \{\text{prox}_{\gamma g}(z) \mid z \in \mathcal{H}, T_{PRS}z = z\}. \quad (3.1.16)$$

That is, if $z^*$ is a fixed point of $T_{PRS}$, then $x^* = x^*_g = x^*_f$ is a solution to Problem 3.1.1, and

$$z^* - x^* = \gamma \nabla g(x^*) \in \gamma \partial g(x^*). \quad (3.1.17)$$

Therefore, the set of fixed points of $T_{PRS}$ is exactly

$$\{x + \gamma w \mid x \in \text{zer}(\partial f + \partial g), w \in (-\partial f(x)) \cap \partial g(x)\}.$$

3.1.10 Fundamental inequalities

Throughout the rest of the chapter we will use the following notation: Every function $f$ is $\mu_f$-strongly convex and $\nabla f$ is $(1/\beta_f)$-Lipschitz. Note that if $\beta_f > 0$, then $f$ is differentiable and $\nabla f = \nabla f$. However, we also allow the strong convexity or Lipschitz differentiability constants to vanish, in which case $\mu_f = 0$ or $\beta_f = 0$ and $f$ may fail to possess either regularity property. Thus, we always assume the inequality holds [11, Theorem 18.15]:

$$f(x) \geq f(y) + \langle x - y, \nabla f(y) \rangle + \max \left\{ \frac{\mu_f}{2} \|x - y\|^2, \frac{\beta_f}{2} \|\nabla f(x) - \nabla f(y)\|^2 \right\}. \quad (3.1.18)$$
In general, $f$ and $g$ can be separable functions on a finite product of Hilbert spaces $\prod_{i=1}^{m} H_i$, i.e. $f(x_1, \cdots, x_m) = \sum_{i=1}^{m} f_i(x_i)$. If smoothness or strong convexity only hold for variables in a subset $I \subseteq \{1, \cdots, m\}$ of this product, we let $\|(x_1, \cdots, x_m)\|_I^2 = \sum_{i \in I} \|x_i\|^2$. Therefore, Equation (3.1.18) motivates the definition of the following nonnegative term:

$$S_f(x, y) = \max \left\{ \mu f \|x - y\|^2, \beta f \|\tilde{\nabla}f(x) - \tilde{\nabla}f(y)\|_I^2 \right\}. \quad (3.1.19)$$

For simplicity we assume that $I = \{1, \cdots, m\}$ for the rest of the chapter.

The following two fundamental inequalities are straightforward modifications of the fundamental inequalities that appeared in Propositions 2.4.1 and 2.4.2 of Chapter 2. When these bounds are iteratively applied, they bound the objective error by the sum of a telescoping sequence and a multiple of the FPR.

**Proposition 3.1.3** (Upper fundamental inequality). Let $z \in \mathcal{H}$, let $z^+ = (T_{PRS})_{\lambda}(z)$, and let $x_f$ and $x_g$ be defined as in Lemma 3.1.2. Then for all $x \in \text{dom}(f) \cap \text{dom}(g)$,

$$4\gamma\lambda(f(x_f) + g(x_g) - f(x) - g(x) + S_f(x_f, x) + S_g(x_g, x)) \leq \|z - x\|^2 - \|z^+ - x\|^2 + \left(1 - \frac{1}{\lambda}\right)\|z^+ - z\|^2. \quad (3.1.20)$$

In our analysis below, we will use the upper inequality

$$4\gamma\lambda(f(x_f) + g(x_g) - f(x^*) - g(x^*) + S_f(x_f, x^*) + S_g(x_g, x^*)) \leq \|z - z^*\|^2 - \|z^+ - x^*\|^2 + 2\langle z - z^*, z^* - x^* \rangle + \left(1 - \frac{1}{\lambda}\right)\|z^+ - z\|^2, \quad (3.1.21)$$

which is obtained from 3.1.20 by letting $x = x^*$ and applying the following identity:

$$\|z - x^*\|^2 - \|z^+ - x^*\|^2 = \|z - z^*\|^2 - \|z^+ - z^*\|^2 + 2\langle z - z^+, z^* - x^* \rangle.$$

**Proposition 3.1.4** (Lower fundamental inequality). Let $z^*$ be a fixed point of $T_{PRS}$, and let $x^* = \text{prox}_{\gamma g}(z^*)$. Then for all $x_f \in \text{dom}(f)$ and $x_g \in \text{dom}(g)$, the lower bound holds:

$$f(x_f) + g(x_g) - f(x^*) - g(x^*) \geq \frac{1}{\gamma} \langle x_g - x_f, z^* - x^* \rangle + S_f(x_f, x^*) + S_g(x_f, x^*). \quad (3.1.22)$$
3.2 Strong convexity

The following theorem will deduce the convergence of $S_f(x^k_f, x^*)$ and $S_g(x^k_g, x^*)$ (see Equation (3.1.19)). In particular, if either $f$ or $g$ is strongly convex and the sequence $(\lambda_j)_{j \geq 0} \subseteq (0, 1]$ is bounded away from zero, then $x^k_f$ and $x^k_g$ converge strongly to a minimizer of $f + g$. Equation (3.2.1) is the main inequality needed to deduce linear convergence of the relaxed PRS algorithm (Section 3.4), and will reappear several times.

**Theorem 3.2.1 (Auxiliary term bound).** Suppose that $(z^j)_{j \geq 0}$ is generated by Algorithm 4. Then for all $k \geq 0$,

$$8\gamma \lambda_k (S_f(x^k_f, x^*) + S_g(x^k_g, x^*)) \quad (3.2.1)$$

$$\leq \|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2 + \left(1 - \frac{1}{\lambda_k}\right) \|z^{k+1} - z^k\|^2. \quad (3.2.2)$$

Therefore, $8\gamma \sum_{i=0}^{\infty} \lambda_k (S_f(x^i_f, x^*) + S_g(x^i_g, x^*)) < \|z^0 - z^*\|^2$, and

1. **Best iterate convergence:** If $\Lambda := \inf_{j \geq 0} \lambda_j > 0$, then

$$\min_{i=0, \ldots, k} \left\{ S_f(x^i_f, x^*) + S_g(x^i_g, x^*) \right\} \leq \frac{\|z^0 - z^*\|^2}{8\gamma \Lambda (k + 1)},$$

and thus

$$\min_{i=0, \ldots, k} \left\{ S_f(x^i_f, x^*) \right\} = o\left(\frac{1}{k + 1}\right) \quad \text{and} \quad \min_{i=0, \ldots, k} \left\{ S_g(x^i_g, x^*) \right\} = o\left(\frac{1}{k + 1}\right).$$

2. **Ergodic convergence:** Let $x^k_f = (1/\Lambda_k) \sum_{i=0}^{k} \lambda_i x^i_f$ and $x^k_g = (1/\Lambda_k) \sum_{i=0}^{k} \lambda_i x^i_g$. Then

$$S_f(x^k_f, x^*) + S_g(x^k_g, x^*) \leq \frac{\|z^0 - z^*\|^2}{8\gamma \Lambda_k},$$

where

$$S_f(x^k_f, x^*) := \max \left\{ \frac{\mu_f}{2} \|\tilde{x}^k_f - x^*\|^2, \frac{\beta_f}{2} \left\| \frac{1}{\Lambda_k} \sum_{i=0}^{k} \tilde{\nabla} f(x^i_f) - \tilde{\nabla} f(x^*) \right\|^2 \right\}$$

and $S_g(x^k_g, x^*)$ is similarly defined.
3. **Nonergodic convergence:** If $\tau = \inf_{j \geq 0} \lambda_j (1 - \lambda_j) > 0$, then

$$S_f(x^k_f, x^*) + S_g(x^k_g, x^*) \leq \frac{\|z^0 - z^*\|^2}{4\gamma \sqrt{\tau(k + 1)}},$$

and thus

$$S_f(x^k_f, x^*) + S_g(x^k_g, x^*) = o \left( \frac{1}{\sqrt{k + 1}} \right).$$

**Proof.** By assumption, the relaxation parameters satisfy $\lambda_k \leq 1$. Therefore, Equation (3.2.1) is a consequence of the following inequalities:

$$8\gamma \lambda_k (S_f(x^k_f, x^*) + S_g(x^k_g, x^*)) \leq 4\gamma \lambda_k (f(x^k_f) + g(x^k_g) - f(x^*) - g(x^*) + S_f(x^k_f, x^*) + S_g(x^k_g, x^*)) - 2\langle z^k - z^{k+1}, z^* - x^* \rangle \leq \|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2 + \left( 1 - \frac{1}{\lambda_k} \right) \|z^{k+1} - z^k\|^2 \leq \|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2. \quad (3.2.3)$$

Part 1 follows from Lemma 3.1.1, and Part 2 follows from Jensen’s inequality applied to $\| \cdot \|^2$. 

Fix $k \geq 0$, let $z_\lambda = (T_{PRS})_\lambda z^k$ for $\lambda \in [0, 1]$, and note that $z_\lambda - z^k = \lambda (T_{PRS} z^k - z^k)$. Proposition 3.1.2 shows that $\|z_\lambda - z^*\| \leq \|z^k - z^*\|$ (Equation (3.1.12)) and that the sequence $(\|z^j - z^*\|)_{j \geq 0}$ is nonincreasing. Therefore, Part 3 is a consequence of the cosine
rule, Proposition 3.1.2, Equation (3.2.1), and the following inequalities:

\[
S_f(x^k_f, x^*) + S_g(x^k_g, x^*) \\
\leq \inf_{\lambda \in [0,1]} \frac{1}{8\gamma \lambda} \left( \|z^k - z^*\|^2 - \|z_\lambda - z^*\|^2 + \left( 1 - \frac{1}{\lambda} \right) \|z^k - z_\lambda\|^2 \right) \\
= \inf_{\lambda \in [0,1]} \frac{1}{8\gamma \lambda} \left( 2\langle z_\lambda - z^*, z^k - z_\lambda \rangle + 2 \left( 1 - \frac{1}{2\lambda} \right) \|z_\lambda - z^k\|^2 \right) \\
\leq \frac{\|z_{1/2} - z^*\| \|z^k - z_{1/2}\|}{2\gamma} \\
\leq \frac{\|z^0 - z^*\| \|z^k - z_{1/2}\|}{2\gamma} \\
\leq \frac{\|z^0 - z^*\|^2}{4\gamma \sqrt{\tau (k + 1)}}.
\]

The little-o convergence rate follows because \( S_f(x^k_f, x^*) + S_g(x^k_g, x^*) \) is bounded by a multiple of the square root of the FPR. \( \square \)

It is not clear whether the “best iterate” convergence results of Theorem 3.2.1 can be improved to a convergence rate for the entire sequence because the values \( S_f(x^k_f, x) \) and \( S_g(x^k_g, x) \) are not necessarily monotonic.

### 3.3 Lipschitz derivatives

In this section, we study the convergence rate of relaxed PRS under the following assumption.

**Assumption 3.3.1.** The gradient of at least one of the functions \( f \) and \( g \) is Lipschitz.

Throughout this section, Lemma 3.1.1 will be used repeatedly to deduce the convergence rates of summable sequences. In general, because we can only deduce the summability and not the monotonicity of the objective errors in Problem 3.1.1, we can only show that the smallest objective error after \( k \) iterations is of order \( o(1/(k + 1)) \). If \( \lambda_k \equiv 1/2 \), the implicit stepsize parameter \( \gamma \) is small enough, and the gradient of \( g \) is \((1/\beta)\)-Lipschitz,
we show that a sequence that dominates the objective error is monotonic and summable, and deduce a convergence rate for the entire sequence.

Because each step of the relaxed PRS algorithm is generated by a proximal operator, it may seem strange that the choice of stepsize $\gamma$ affects the convergence rate of relaxed PRS. This is certainly not the case for the proximal point algorithm, which achieves an $o(1/(k+1))$ convergence rate by Theorem 2.3.3 of Chapter 2. A possible explanation is that the reflection operator of a differentiable function is the composition of averaged operators

$$\text{refl}_{\gamma g} = (I - \gamma \nabla g) \circ \text{prox}_{\gamma g}$$

whenever $\gamma < 2\beta$, and, therefore, it is averaged [11, Propositions 4.32 and 4.33]. Thus, although $T_{\text{PRS}}$ is not necessarily averaged when $f$ or $g$ is differentiable, the individual reflection operators enjoy a stronger contraction property [11, Proposition 4.25] as long as $\gamma$ is small enough. As soon as $\gamma$ is too large, we seem to lose monotonicity of various sequences that arise in our analysis.

### 3.3.1 The general case: best iterate convergence rate

The following Theorem will be used several times throughout our analysis.

**Theorem 3.3.1** (Descent theorem/Baillon-Haddad). Suppose that $g : \mathcal{H} \rightarrow (-\infty, \infty]$ is closed, proper, convex, and differentiable. If $\nabla g$ is $(1/\beta)$-Lipschitz, then for all $x, y \in \mathcal{H}$, we have the upper bound

$$g(x) \leq g(y) + \langle x - y, \nabla g(y) \rangle + \frac{1}{2\beta} \|x - y\|^2, \quad (3.3.1)$$

and the cocoercive inequality

$$\beta \|\nabla g(x) - \nabla g(y)\|^2 \leq \langle x - y, \nabla g(x) - \nabla g(y) \rangle. \quad (3.3.2)$$

**Proof.** See [11, Theorem 18.15(iii)] for Equation (3.3.1), and [3] for Equation (3.3.2).
The next proposition bounds the objective error by a summable sequence.

**Proposition 3.3.1** (Fundamental inequality under Lipschitz assumptions). Let \( z \in \mathcal{H} \), let \( z^+ = (T_{PRS})_\lambda z \), let \( z^* \) be a fixed point of \( T_{PRS} \), and let \( x^* = \text{prox}_{\gamma g}(z^*) \). If \( \nabla f \) (respectively \( \nabla g \)) is \((1/\beta)\)-Lipschitz, then for \( x = x_g \) (respectively \( x = x_f \)),

\[
4 \gamma \lambda \left( f(x) + g(x) - f(x^*) - g(x^*) \right)
\leq \begin{cases} 
\|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left( 1 + \frac{1}{2\lambda \left( \frac{\gamma}{\beta} - 1 \right) \beta} \right) \|z - z^+\|^2, & \text{if } \gamma \leq \beta; \\
\left( 1 + \frac{(\gamma - \beta)}{2\beta} \right) \left( \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \|z - z^+\|^2 \right), & \text{otherwise.}
\end{cases}
\]

**Proof.** Because \( \nabla f \) is \((1/\beta)\)-Lipschitz, we have

\[
f(x_g) \overset{(3.3.1)}{\leq} f(x_f) + \langle x_g - x_f, \nabla f(x_f) \rangle + \frac{1}{2\beta} \|x_g - x_f\|^2,
\]

\[
S_f(x_f, x^*) \overset{(3.1.18)}{\geq} \frac{\beta}{2} \|\nabla f(x_f) - \nabla f(x^*)\|^2.
\]

We now derive some identities that will be used below to bound \( f(x_g) + g(x_g) - f(x^*) - g(x^*) \). By applying the identity \( z^* - x^* = \gamma \widetilde{\nabla} g(x^*) = -\gamma \nabla f(x^*) \) (Equation (3.1.17)), the cosine rule (3.1.5), and Equation (3.1.15) multiple times, we have

\[
2 \langle z - z^+, z^* - x^* \rangle + 4 \gamma \lambda \langle x_g - x_f, \nabla f(x_f) \rangle
= 4 \gamma \lambda \langle x_g - x_f, \nabla f(x_f) - \nabla f(x^*) \rangle
= 4 \lambda \langle \gamma \widetilde{\nabla} g(x_g) + \gamma \nabla f(x_f), \gamma \nabla f(x_f) - \gamma \nabla f(x^*) \rangle
= 2 \lambda \left( \|x_f - x_g\|^2 + \|\gamma \nabla f(x_f) - \gamma \nabla f(x^*)\|^2 \right)
- \|\gamma \widetilde{\nabla} g(x_g) - \gamma \widetilde{\nabla} g(x^*)\|^2.
\]

By Equation (3.1.15)

\[
\left( 1 - \frac{1}{\lambda} \right) \|z - z^+\|^2 + 2 \lambda \left( \frac{\gamma}{\beta} + 1 \right) \|x_g - x_f\|^2 = \left( 1 + \frac{(\gamma - \beta)}{2\beta \lambda} \right) \|z - z^+\|^2.
\]
Using the above two identities, we have

\[ 4\gamma\lambda(f(x_g) + g(x_g) - f(x^*) - g(x^*)) \]
\[ \leq 4\gamma\lambda(f(x_f) + g(x_g) - f(x^*) - g(x^*)) + 4\gamma\lambda(x_g - x_f, \nabla f(x_f)) + \frac{2\gamma\lambda}{\beta} ||x_g - x_f||^2 \]
\[ \leq ||z - z^*||^2 - ||z^+ - z^*||^2 + \left( 1 - \frac{1}{\lambda} \right)||z - z^+||^2 + 2\gamma\lambda \left( \frac{\gamma}{\beta} + 1 \right)||x_g - x_f||^2 \]
\[ + 2\gamma\lambda||\nabla f(x_f) - \nabla f(x^*)||^2 - 4\gamma\lambda S_f(x_f, x^*) - 2\lambda||\nabla g(x_g) - \nabla g(x^*)||^2 \]
\[ \leq ||z - z^*||^2 - ||z^+ - z^*||^2 + \left( 1 - \frac{1}{\lambda} \right)||z - z^+||^2 + \left( 1 + \frac{(\gamma - \beta)}{2\beta\lambda} \right)||z - z^+||^2 \]
\[ + 2\gamma\lambda(\gamma - \beta)||\nabla f(x_f) - \nabla f(x^*)||^2. \]

If \( \gamma \leq \beta \), we can drop the last term. If \( \gamma > \beta \), we apply the upper bound on \( S_f(x_f, x) \) in (3.2.1) to get

\[ 2\gamma\lambda(\gamma - \beta)||\nabla f(x_f) - \nabla f(x^*)||^2 \]
\[ \leq \frac{(\gamma - \beta)}{2\beta} \left( ||z - z^*||^2 - ||z^+ - z^*||^2 + \left( 1 - \frac{1}{\lambda} \right)||z - z^+||^2 \right), \]

and the result follows.

If \( \nabla g \) is \((1/\beta)\)-Lipschitz, the argument is symmetric, so we omit the proof. \( \square \)

Proposition 3.3.1 shows that the the objective error is summable whenever \( f \) or \( g \) is Lipschitz and \((\lambda_j)_{j \geq 0}\) is chosen properly. A direct application of Lemma 3.1.1 yields a convergence rate for the objective error. Depending on the choice of \( \gamma \) and \((\lambda_j)_{j \geq 0}\), we can achieve several different rates. In the following Theorem we only analyze a few such choices.

**Theorem 3.3.2** (Best iterate convergence under Lipschitz assumptions). Let \( z \in \mathcal{H} \), let \( z^+ = (T_{PRS})_\lambda z \), let \( z^* \) be a fixed point of \( T_{PRS} \), and let \( x^* = \text{prox}_{\gamma g}(z^*) \). Suppose that
\[ \tau = \inf_{j \geq 0} \lambda_j (1 - \lambda_j) > 0, \text{ and let } \lambda = \inf_{j \geq 0} \lambda_j. \text{ If } \nabla f \text{ (respectively } \nabla g) \text{ is } (1/\beta)\text{-Lipschitz, and } x^k = x^k_f \text{ (respectively } x^k = x^k_g), \text{ then} \]

\[ \min_{i=0 \ldots k} \left\{ f(x^i) + g(x^i) - f(x^*) - g(x^*) \right\} = o \left( \frac{1}{k+1} \right). \]

**Proof.** Proposition 3.1.2 proves the following bound:

\[ \inf_{j \geq 0} \frac{1 - \lambda_j}{\lambda_j} \sum_{i=0}^{\infty} \| z^k - z^{k+1} \|^2 \leq \sum_{i=0}^{\infty} \tau_i \| T_{PRS} z^i - z^i \|^2 \leq \| z^0 - z^* \|^2. \]

Therefore, the proof follows from Part 4 of Lemma 3.1.1 applied to the summable upper bound in Proposition 3.3.1, which bounds the objective error. Note that under different choices of \((\lambda_j)_{j \geq 0}\) and \(\gamma\), we get the bounds:

\[ \min_{i=0 \ldots k} \left\{ f(x^i) + g(x^i) - f(x^*) - g(x^*) \right\} \leq \frac{\| z^0 - z^* \|^2}{4\gamma \lambda (k+1)} \times \begin{cases} 1, & \text{if } \gamma \leq \beta \text{ and } (\lambda_j)_{j \geq 0} \subseteq \left[ \frac{\lambda}{2}, \frac{1}{2} \right); \\ 1 + 1/\left( \inf_{j \geq 0} \frac{1 - \lambda_j}{\lambda_j} \right), & \text{if } \gamma \leq \beta; \\ \left( 1 + \frac{(\gamma - \beta)}{2\beta} \right) \left( 1 + 1/\left( \inf_{j \geq 0} \frac{1 - \lambda_j}{\lambda_j} \right) \right), & \text{otherwise.} \end{cases} \]

The main conclusion of Theorem 3.3.2 is that as long as \(\tau > 0\),

the “best” relaxed PRS iterate converges with rate \(o(1/(k+1))\) for any input parameters.

This result should be compared with the known convergence properties of the FBS algorithm, which has order \(o(1/(k+1))\) for small \(\gamma\), but may even fail to converge if \(\gamma\) is too large. See Section 3.1.5 for more on the distinction between FBS and relaxed PRS.
3.3.2 Constant relaxation and better rates

In this section, we study the convergence rate of DRS under the assumption

**Assumption 3.3.2.** The function $g$ is differentiable on $\text{dom}(f) \cap \text{dom}(g)$, the gradient $\nabla g$ is $(1/\beta)$-Lipschitz, and the sequence of relaxation parameters $(\lambda_j)_{j \geq 0}$ is constant and equal to $1/2$.

With these assumptions, we will show that for a special choice of $\theta^*$ (Lemma 3.3.2) and for $\gamma$ small enough, the following sequence is monotonic and summable (Propositions 3.3.3 and 3.3.5):

$$
\left( 2\gamma (f(x_j) + g(x_j) - f(x) - g(x)) \\
+ \theta^* \gamma^2 \|\nabla g(x_j+1) - \nabla g(x_j)\|^2 + \frac{(1 - \theta^*) \gamma^2}{\beta^2} \|x_j+1 - x_j\|^2 \right)_{j \geq 0}.
$$

(3.3.5)

We then use Lemma 3.1.1 to deduce $f(x_j) + g(x_j) - f(x) - g(x) = o(1/(k + 1))$.

There are several other simpler monotonic and summable sequences that dominate the objective error. For example, if we choose $\theta^* = 1$, we can drop the last term in Equation (3.3.5), but we can no longer use this sequence to help deduce the convergence rate of the FPR in Theorem 3.3.4. Thus, we choose to analyze the slightly complicated sequence in Equation (3.3.5) in order to provide a unified analysis for all results in this section.

The following two results are well known, but we include some of the proofs for completeness. They will help us tighten the bounds that we develop below.

**Lemma 3.3.1** (Extra contraction of derivative operator). Suppose that $\nabla g$ is $(1/\beta)$-Lipschitz, and let $x, y \in \mathcal{H}$. If $x^+ = \text{prox}_{\gamma g}(x)$ and $y^+ = \text{prox}_{\gamma f}(y)$, then

$$
\|\nabla g(x^+) - \nabla g(y^+)\|^2 \leq \frac{1}{\gamma^2 + \beta^2} \|x - y\|^2.
$$

(3.3.6)

**Proof.** From the identity $\gamma \nabla g(x^+) = x - x^+$, the contraction property in Corollary 2, and
the Lipschitz continuity of $\nabla g$ we have
\begin{align*}
\beta^2 \| \nabla g(x^+) - \nabla g(y^+) \|^2 &\leq \| x^+ - y^+ \|^2; \\
\gamma^2 \| \nabla g(x^+) - \nabla g(y^+) \|^2 &\leq \| x - y \|^2 - \| x^+ - y^+ \|^2
\end{align*}

Adding both equations and rearranging proves the result. \hfill \square

The following is a direct corollary of the descent theorem (Theorem 3.3.1).

**Corollary 3.3.1** (Joint descent theorem). If $g$ is differentiable and $\nabla g$ is $(1/\beta)$-Lipschitz, then for all pairs $x, y \in \text{dom}(f)$, points $z \mathcal{H}$, and subgradients $\tilde{\nabla} f(x) \in \partial f(x)$, we have
\begin{equation}
f(x) + g(x) \leq f(y) + g(y) + \langle x - y, \nabla g(z) + \tilde{\nabla} f(x) \rangle + \frac{1}{2\beta} \| z - x \|^2. \tag{3.3.7}
\end{equation}

**Proof.** Inequality (3.3.7) follows from adding the upper bound
\begin{align*}
g(x) - g(y) &\leq g(z) - g(y) + \langle x - z, \nabla g(z) \rangle + \frac{1}{2\beta} \| z - x \|^2 \\
&\leq \langle x - y, \nabla g(z) \rangle + \frac{1}{2\beta} \| z - x \|^2,
\end{align*}
with the subgradient inequality: $f(x) \leq f(y) + \langle x - y, \tilde{\nabla} f(x) \rangle$. \hfill \square

The following theorem develops an alternative fundamental inequality to the one presented in Proposition 3.3.1.

**Proposition 3.3.2** (Fundamental inequality for differentiable functions). For all $x \in \text{dom}(f)$,
\begin{align*}
2\gamma (f(x^k_f) + g(x^k_f) - f(x) - g(x)) &+ \left( 2\gamma \beta - \frac{\gamma^3}{\beta} \right) \| \nabla g(x_g^{k+1}) - \nabla g(x_g^k) \|^2 \\
+ \| x_g^{k+1} - x \|^2 + \| x_g^{k+1} - x_g^k \|^2 &\leq \| x_g^k - x \|^2. \tag{3.3.8}
\end{align*}
Proof. The following identities are straightforward from Lemma 3.1.2:

\[ x^k_g - x^{k+1}_g = \gamma(\nabla g(x^{k+1}_g + \tilde{\nabla} f(x^k_f)) \quad \text{and} \quad x^k_f - x^{k+1}_g = \gamma(\nabla g(x^{k+1}_g) - \nabla g(x^k_g)). \tag{3.3.9} \]

Therefore,

\[
2\gamma(f(x^k_f) + g(x^k_f) - f(x) - g(x)) + \left(2\gamma \beta - \frac{\gamma^3}{\beta}\right) \| \nabla g(x^{k+1}_g) - \nabla g(x^k_g) \|^2 \\
\leq 2\gamma(x^k_f - x, \tilde{\nabla} f(x^k_f) + \nabla g(x^{k+1}_g)) + \frac{\gamma}{\beta} \| x^k_f - x^{k+1}_g \|^2 \\
+ \left(2\gamma \beta - \frac{\gamma^3}{\beta}\right) \| \nabla g(x^{k+1}_g) - \nabla g(x^k_g) \|^2 \\
\overset{(3.3.7)}{=} 2(x^k_f - x, x^k_g - x^{k+1}_g) + 2\gamma \beta \| \nabla g(x^{k+1}_g) - \nabla g(x^k_g) \|^2 \\
= 2\gamma(x^{k+1}_g - x, x^k_g - x^{k+1}_g) + 2(x^k_f - x^{k+1}_g, x^k_g - x^{k+1}_g) \\
+ 2\gamma \beta \| \nabla g(x^{k+1}_g) - \nabla g(x^k_g) \|^2 \\
\overset{(3.3.9)}{\leq} \| x^k_g - x \|^2 - \| x^{k+1}_g - x \|^2 - \| x^k_g - x^{k+1}_g \|^2 \\
+ 2\gamma \| \nabla g(x^{k+1}_g) - \nabla g(x^k_g) \|^2 = 2\gamma \beta \| \nabla g(x^{k+1}_g) - \nabla g(x^k_g) \|^2 \\
\overset{(3.3.2)}{\leq} \| x^k_g - x \|^2 - \| x^{k+1}_g - x \|^2 - \| x^k_g - x^{k+1}_g \|^2. \tag{3.3.10} \\
\]

Equation (3.3.8) now follows by rearranging Equation (3.3.10). \qed

The following proposition uses the fundamental inequality in Proposition 3.3.2 evaluated at the point \( x = x^{k-1}_f \) to construct a monotonic sequence that dominates the objective error. We introduce a factor \( \theta \in [0, 1] \) that we will optimize in Lemma 3.3.2 in order to maximize the range of \( \gamma \) for which the sequence remains monotonic.

Proposition 3.3.3 (Monotonicity). For scalars \( \theta \in [0, 1] \) and integers \( k \geq 1 \), the following bound holds:

\[
2\gamma(f(x^k_f) + g(x^k_f) - f(x^*) - g(x^*)) + \left(2\gamma \beta - \frac{\gamma^3}{\beta}\right) \| \nabla g(x^{k+1}_g) - \nabla g(x^k_g) \|^2 + \| x^{k+1}_g - x^k_g \|^2 \\
\leq 2\gamma(f(x^{k-1}_f) + g(x^{k-1}_f) - f(x^*) - g(x^*)) + \theta \| \nabla g(x^k_g) - \nabla g(x^{k-1}_g) \|^2 \\
+ \frac{(1 - \theta)\gamma^2}{\beta^2} \| x^k_g - x^{k-1}_g \|^2. \tag{3.3.11} \\
\]
Proof. Plug \( x = x_f^{k-1} \) into Equation (3.3.8) and subtract \( f(x^*) + g(x^*) \) from both sides. Equation (3.3.11) follows from the identity
\[
x_g^k - x_f^{k-1} = \gamma (\nabla g(x_g^{k-1}) - \nabla g(x_g^k)),
\]
the bound \( \|\nabla g(x_g^k) - \nabla g(x_g^{k-1})\| \leq (1/\beta^2)\|x_g^k - x_g^{k-1}\|^2 \), rearranging, and dropping the positive term \( \|x_g^{k+1} - x_f^{k-1}\|^2 \).

We now choose the factor \( \theta \) in order to maximize the range of implicit stepsize parameters \( \gamma \) for which the sequence constructed in Proposition 3.3.3 remains monotonic. 

Lemma 3.3.2 (Maximizing \( \gamma \) range). Let \( \beta > 0 \), and let
\[
\kappa := \sup \left\{ \frac{\gamma}{\beta} \mid \gamma > 0, \theta \in [0, 1], \theta \gamma^2 \leq \left( 2\gamma \beta - \frac{\gamma^3}{\beta} \right), \frac{(1-\theta)\gamma^2}{\beta^2} \leq 1 \right\}. \tag{3.3.12}
\]
Then \( \kappa \) is the positive root of \( x^3 + x^2 - 2x - 1 \). Therefore, \((\gamma^*, \theta^*) = (\kappa \beta, 1 - 1/\kappa^2)\).

Proof. Observe that the constraints on \( \theta \) and \( \gamma \) are equivalent to following inequalities:
\[
1 + \frac{2\gamma}{\beta} - \frac{\gamma^2}{\beta^2} - \frac{\gamma^3}{\beta^3} \geq (\theta - 1)\frac{\gamma^2}{\beta^2} + 1 \geq 0. \tag{3.3.13}
\]
The left hand side of Equation (3.3.13) is monotonically decreasing in \( \gamma \) for all \( \gamma \geq \beta \). Furthermore, if \( \gamma = \kappa \beta \), then the left hand side is 0. Thus, \( \gamma^* \leq \kappa \beta \). Finally, for every \( \gamma \in [0, \kappa \beta] \), the scalar \( \theta^* = 1 - \beta^2/\gamma^2 \) satisfies \( (\theta - 1)(\gamma^2/\beta^2) + 1 \geq 0 \). Therefore, \((\gamma^*, \theta^*) = (\kappa \beta, 1 - 1/\kappa^2)\). \(\square\)

Remark 3.3.1. Throughout the rest of the chapter, we will let \( \kappa = 1/\sqrt{1-\theta^*} \approx 1.24698 \) where \( \theta^* \) is defined in Lemma 3.3.2. Note that the inequality constraints in Equation (3.3.12) become equalities for the pair \((\gamma^*, \theta^*)\).

We will need the following bound in several of the proofs below.

Proposition 3.3.4 (Gradient sum bound). For all \( \gamma > 0 \)
\[
\sum_{i=0}^{\infty} \|\nabla g(x_g^i) - \nabla g(x_g^{i+1})\|^2 \leq \frac{1}{\gamma^2 + \beta^2} \|z^0 - z^*\|^2. \tag{3.3.14}
\]
Proof. From Lemma 3.3.1 and the Fejér type in equality in Equation (3.1.12):

\[
\|\nabla g(x^k_g) - \nabla g(x^{k+1}_g)\|^2 \leq \left(1/(\gamma^2 + \beta^2)\right)\|z^k - z^{k+1}\|^2 \\
\leq \left(1/(\gamma^2 + \beta^2)\right)\left(\|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2\right).
\] (3.3.15)

Therefore, the result follows by summing (3.3.15). \qed

The following proposition computes an upper bound of the sum of the sequence in Equation (3.3.11).

**Proposition 3.3.5** (Summability). If \(\gamma < \kappa \beta\), choose \(\theta = \theta^*\) as in Lemma 3.3.2; otherwise, set \(\theta = 1\). Then

\[
\sum_{i=0}^{\infty} \left(2\gamma(f(x^k_f) + g(x^k_g) - f(x^*) - g(x^*)) + \theta\gamma^2\|\nabla g(x^{k+1}_g) - \nabla g(x^k_g)\|^2 \\
+ \left(1 - \theta\right)\gamma^2\|x^{k+1}_g - x^k_g\|^2 \right) \\
\leq \left\{ \begin{array}{ll}
\|x^0_g - x^*\|^2, & \text{if } \gamma < \kappa \beta; \\
\|x^0_g - x^*\|^2 + \frac{1}{\beta\gamma^2} \left(\frac{\gamma^3}{\beta} - 2\gamma \beta + \gamma^2 - \beta^2\right)\|z^0 - z^*\|^2, & \text{otherwise}.
\end{array} \right.
\] (3.3.16)

Proof. First note that:

\[-\|x^k_g - x^{k+1}_g\|^2 \leq -\beta^2\|\nabla g(x^k_g) - \nabla g(x^{k+1}_g)\|^2.\]

In addition, for either choice of \(\theta\) we have \((1-\theta)\gamma^2/\beta^2 - 1 \leq 0\). Thus, from Equation (3.3.8)

\[
2\gamma(f(x^k_f) + g(x^k_g) - f(x^*) - g(x^*)) + \theta\gamma^2\|\nabla g(x^{k+1}_g) - \nabla g(x^k_g)\|^2 \\
+ \left(1 - \theta\right)\gamma^2\|x^{k+1}_g - x^k_g\|^2 \\
\leq \|x^k_g - x^*\|^2 - \|x^{k+1}_g - x^*\|^2 \\
+ \left(\frac{\gamma^3}{\beta} - 2\gamma \beta + \theta\gamma^2\right)\|\nabla g(x^{k+1}_g) - \nabla g(x^k_g)\|^2 + \left(\frac{(1 - \theta)\gamma^2}{\beta^2} - 1\right)\|x^k_g - x^{k+1}_g\|^2 \\
\leq \|x^k_g - x^*\|^2 - \|x^{k+1}_g - x^*\|^2 + \left(\frac{\gamma^3}{\beta} - 2\gamma \beta + \gamma^2 - \beta^2\right)\|\nabla g(x^{k+1}_g) - \nabla g(x^k_g)\|^2.
\] (3.3.17)
The last line of Equation (3.3.17) is negative if, and only if, $\gamma \leq \kappa \beta$. This proves the first bound in Equation (3.3.16). The second bound follows from the sum bound in Equation (3.3.14).

We are now ready to deduce the objective error convergence rate for the DRS algorithm when $\nabla g$ is Lipschitz. Our bounds show that

DRS is at least as fast as FBS whenever $\gamma$ is small enough.

Additionally, we show that the convergence rate of the best iterate has essentially the same constant for a large range of $\gamma$. When $\gamma$ is large, the best iterate still enjoys the convergence rate $o(1/(k+1))$, albeit with a larger constant (Theorem 3.3.2). The rates we derive are the best possible for this algorithm, as shown by Theorem 2.7.3 of Chapter 2.

**Theorem 3.3.3** (Differentiable function convergence rate). Let $\rho \approx 2.2056$ be the positive real root of $x^3 - 2x^2 - 1$. Then

$$
\min_{i=0,\ldots,k} \{ f(x^i_f) + g(x^i_f) - f(x^*) - g(x^*) \} \leq \frac{1}{2\gamma(k+1)} \begin{cases} 
\|x_0^f - x^*\|^2, & \text{if } \gamma < \rho \beta; \\
\|x_0^f - x^*\|^2 + \frac{1}{\beta^2 + \gamma} \left( \frac{x_0^f}{\beta} - 2\gamma \beta - \beta^2 \right) \|z_0^f - z^*\|^2, & \text{otherwise}; 
\end{cases}
$$

and

$$
f(x_{f}^{k_{\text{best}}}^f) + g(x_{f}^{k_{\text{best}}}^f) - f(x^*) - g(x^*) = o \left( \frac{1}{k+1} \right).$$

Furthermore, if $\gamma < \kappa \beta$, then

$$
f(x_f^k) + g(x_f^k) - f(x^*) - g(x^*) \leq \frac{\|x_0^f - x^*\|^2}{2\gamma(k+1)};$$

$$f(x_f^k) + g(x_f^k) - f(x^*) - g(x^*) = o \left( \frac{1}{k+1} \right).$$
Proof. To prove the “k_{best}” bounds, rearrange the upper inequality in Equation (3.3.8) to

\[
2\gamma(f(x_{f}^{k}) + g(x_{f}^{k}) - f(x^{*}) - g(x^{*}))
\leq \|x_{f}^{k} - x^{*}\| - \|x_{f}^{k+1} - x^{*}\|^2 + \left(\frac{\gamma^3}{\beta} - 2\gamma\beta\right) \|\nabla g(x_{g}^{k+1}) - \nabla g(x_{g}^{k})\|^2 - \|x_{g}^{k} - x_{g}^{k+1}\|^2
\leq \|x_{g}^{k} - x^{*}\| - \|x_{g}^{k+1} - x^{*}\|^2 + \left(\frac{\gamma^3}{\beta} - 2\gamma\beta - \beta^2\right) \|\nabla g(x_{g}^{k+1}) - \nabla g(x_{g}^{k})\|^2,
\]

(3.3.18)

where the last line follows from the bound \(-\|x_{g}^{k} - x_{g}^{k+1}\|^2 \leq -\beta^2 \|\nabla g(x_{g}^{k}) - \nabla g(x_{g}^{k+1})\|^2\).

Note that \(\frac{\gamma^3}{\beta} - 2\gamma\beta - \beta^2 \leq 0\) if, and only if, \(\gamma \leq \rho\beta\) where \(\rho\) is the positive root of \(x^3 - 2x^2 - 1\). Therefore, the result follows by summing Equation (3.3.18) and applying Part 4 of Lemma 3.1.1.

If \(\gamma \leq \kappa\beta\), then \((\frac{\gamma^3}{\beta} - 2\gamma\beta + \theta^*\gamma^2) \leq 0\) and \((1 - \theta^*)\gamma^2 / \beta^2 \leq 1\). Therefore, Equation (3.3.11) shows that the sequence

\[
\left(2\gamma(f(x_{f}^{j}) + g(x_{f}^{j}) - f(x) - g(x)) + \theta^*\gamma^2 \|\nabla g(x_{g}^{j+1}) - \nabla g(x_{g}^{j})\|^2 + \frac{(1 - \theta^*)\gamma^2}{\beta^2} \|x_{g}^{j+1} - x_{g}^{j}\|^2\right)_{j \geq 0}
\]

is monotonic. In addition, Equation (3.3.16) shows the sum of this sequence is bounded by \(\|x_{g}^{0} - x^{*}\|^2\). Therefore, the result follows by Lemma 3.1.1.

In Theorem 2.3.3 of Chapter 2, we showed that the FPR convergence rate for the FBS algorithm is \(o(1/(k + 1)^2))\). We complement this result by showing the same is true for DRS whenever \(\gamma\) is small enough. This rate is optimal by Theorem 2.7.3 of Chapter 2.

**Theorem 3.3.4** (Differentiable function FPR rate). Suppose that \(\gamma < \kappa\beta\). Then for all \(k \geq 1\), we have

\[
\|z^{k} - z^{k+1}\|^2 \leq \frac{\beta^2 \|x_{g}^{0} - x^{*}\|^2}{k^2(1 + \gamma / \beta)^2 (\beta^2 - \gamma^2 / \kappa^2)} \quad \text{and} \quad \|z^{k} - z^{k+1}\|^2 = o\left(\frac{1}{k^2}\right).
\]

(3.3.19)

Proof. For all \(k \geq 1\), let

\[
\eta = 1 - \frac{(1 - \theta^*)\gamma^2}{\beta^2} = \frac{\beta^2 + (\theta^* - 1)\gamma^2}{\beta^2} = \frac{\beta^2 - \gamma^2 / \kappa^2}{\beta^2},
\]

93
let \( a_{k-1} = \left( \eta/(1 + \gamma/\beta)^2 \right) \| z^{k+1} - z^k \|^2 \), and let

\[
b_{k-1} = 2\gamma(f(x^k_j) + g(x^k_j) - f(x^*) - g(x^*))
+ \theta^* \gamma^2 \| \nabla g(x^{k+1}_g) - \nabla g(x^k_g) \|^2
+ \frac{(1 - \theta^*) \gamma^2}{\beta^2} \| x^{k+1}_g - x^k_g \|^2.
\]

Because \( z^k = x^k_g + \gamma \nabla g(x^k_g) \) and \( \nabla g \) is \((1/\beta)\)-Lipschitz, we get

\[
\eta \| z^k - z^{k+1} \|^2 \leq \eta \left( 1 + \frac{\gamma}{\beta} \right)^2 \| x^k_g - x^{k+1}_g \|^2.
\]

Therefore, Equation (3.3.11) shows that for all \( k \geq 1 \),

\[
a_{k-1} \leq \eta \| x^k_g - x^{k+1}_g \|^2 \leq b_{k-1} - b_k.
\]

Part 3 of Lemma 3.1.1 applied to the sequences \((a_j)_{j\geq 0}\) and \((b_j)_{j\geq 0}\) with weighting parameters \( \lambda_k \equiv 1 \), (not to be confused with the constant relaxation parameter of the relaxed PRS algorithm), yields

\[
\sum_{i=0}^{\infty} (i + 1) a_i \leq \sum_{i=0}^{\infty} b_i \overset{(3.3.16)}{\leq} \| x^0_g - x^* \|^2.
\]

Part 2 of Proposition 2 shows that \((a_j)_{j\geq 0}\) is monotonic. Therefore, the result follows from Part 1c of Lemma 3.1.1.

\[\square\]

**Remark 3.3.2.** Note that by Theorem 2.3.3 of Chapter 2, the FBS algorithm achieves \( o(1/(k + 1)) \) objective error rate and \( o(1/(k + 1)^2) \) FPR rate as long as \( \gamma < 2\beta \). For the DRS algorithm, our analysis only covers the smaller range \( \gamma \leq \kappa \beta \). It is an open question whether \( \kappa \) can be improved for the DRS algorithm.

### 3.4 Linear convergence

In this section, we study the convergence rate of relaxed PRS under the assumption

**Assumption 3.4.1.** The gradient of at least one of the functions \( f \) and \( g \) is Lipschitz, and at least one of the functions \( f \) and \( g \) is strongly convex. In symbols: \( (\mu_f + \mu_g)(\beta_f + \beta_g) > 0 \).
Linear convergence of relaxed PRS is expected whenever Assumption 3.4.1 is true. In addition, by the strong convexity of $f + g$, the minimizer of Problem (3.1.1) is unique.

The following proposition lists some consequences of linear convergence of the relaxed PRS sequence $(z^j)_{j \geq 0}$.

**Proposition 3.4.1** (Consequences of linear convergence). Let $(C_j)_{j \geq 0} \subseteq [0, 1]$ be a positive scalar sequence, and suppose that for all $k \geq 0$,

$$\|z^{k+1} - z^*\| \leq C_k \|z^k - z^*\|.$$  \hfill (3.4.1)

Fix $k \geq 1$. Then

$$\|x^k_g - x^*\|^2 + \gamma^2 \|\tilde{\nabla} g(x^k_g) - \tilde{\nabla} g(x^*)\|^2 \leq \|z^0 - z^*\|^2 \prod_{i=0}^{k-1} C_i^2;$$

$$\|x^k_f - x^*\|^2 + \gamma^2 \|\tilde{\nabla} f(x^k_f) - \tilde{\nabla} f(x^*)\|^2 \leq \|z^0 - z^*\|^2 \prod_{i=0}^{k-1} C_i^2.$$

If $\lambda < 1$, then the FPR rate holds:

$$\|(T_{PRS})_\lambda z^k - z^k\| \leq \sqrt{\frac{\lambda}{1 - \lambda}} \|z^0 - z^*\| \prod_{i=0}^{k-1} C_i.$$

Consequently, if the gradient $\nabla f$ (respectively $\nabla g$), is $(1/\beta)$-Lipschitz and $x^k = x^k_g$ (respectively $x^k = x^k_f$), then

$$f(x^k) + g(x^k) - f(x^*) - g(x^*) \leq \frac{\|z^0 - z^*\|^2}{\gamma} \prod_{i=0}^{k-1} C_i^2 \times \begin{cases} 1, & \text{if } \gamma \leq \beta; \\ 1 + \frac{(\gamma - \beta)}{2\beta}, & \text{otherwise.} \end{cases}$$

**Proof.** The bounds for $x^k_g$ and $x^k_f$ follow because

$$\|x^k_g - x^*\|^2 + \gamma^2 \|\nabla g(x^k_g) - \nabla g(x^*)\|^2 \leq \|z^k - z^*\|^2,$$

and

$$\|x^k_f - x^*\|^2 + \gamma^2 \|\tilde{\nabla} f(x^k_f) - \tilde{\nabla} f(x^*)\|^2 \leq \|\text{refl}_{\gamma_f}(z^k) - \text{refl}_{\gamma_f}(z^*)\|^2 \leq \|z^k - z^*\|^2.$$

by Part 2 of Proposition 3.1.1, the nonexpansiveness of $\text{refl}_{\gamma_f}$, and Equation (3.6.13).
The FPR convergence rate follows from the Fejér-type inequality in Equation (3.1.12).

Now fix \( k \geq 1 \), and let \( z_{\lambda} = (T_{\text{PRS}})_{\lambda} z^k \) for all \( \lambda \in [0, 1] \). Then Proposition 3.3.1 shows that:

\[
f(x^k) + g(x^k) - f(x^*) - g(x^*)
\leq \inf_{\lambda \in [0, 1]} \frac{1}{4\gamma \lambda} \left\{ \| z^k - z^* \|^2 - \| z\lambda - z^* \|^2 + \left( 1 + \frac{1}{2\lambda} \left( \frac{\gamma}{\beta} - 1 \right) \right) \| z^k - z\lambda \|^2, \right.
\]

\[
\text{if } \gamma \leq \beta;
\]

\[
\left( 1 + \frac{\gamma - \beta}{2\beta} \right) \left( \| z^k - z* \|^2 - \| z\lambda - z^* \|^2 + \| z^k - z\lambda \|^2 \right), \quad \text{otherwise.}
\]

The objective error rate now follows from Equation (3.4.2) and the FPR convergence rate.

Whenever \( \sup_{j \geq 0} C_j < 1 \), Proposition 3.4.1 gives the linear convergence rates of the sequences \((z^j)_{j \geq 0}\), \((x^j_g)_{j \geq 0}\) and \((x^j_f)_{j \geq 0}\), the subgradient error, the FPR, and the objective error. In the following sections, we will prove Inequality (3.6.13) holds under several different regularity assumptions on \( f \) and \( g \). In each case we leave it to the reader to apply Proposition 3.4.1.

### 3.4.1 Solely regular \( f \) or \( g \)

Throughout this subsection, at least one of the functions \( f \) and \( g \) will carry both regularity properties. In symbols: \( \mu_f \beta_f + \mu_g \beta_g > 0 \).

The following theorem recovers [85, Proposition 4] as a special case (\( \lambda_k = 1/2 \)).

**Theorem 3.4.1** (Linear convergence with regularity of \( g \)). Let \( z^* \) be a fixed point of \( T_{\text{PRS}} \), let \( x^* = \text{prox}_{\gamma g}(z^*) \), and suppose that \( \mu_g \beta_g > 0 \). For all \( \lambda \in [0, 1] \), let \( C(\lambda) := (1 - 4\gamma \lambda \mu_g/(1 + \gamma/\beta_g)^2)^{1/2} \). Then for all \( k \geq 0 \),

\[
\| z^{k+1} - z^* \| \leq C(\lambda_k) \| z^k - z^* \|. \tag{3.4.3}
\]
Proof. Theorem 3.2.1 bounds the distance of $x^k_g$ to the minimizer

$$\frac{8\gamma \lambda_k \mu_g}{2} \| x^k_g - x^* \|^2 \overset{(3.2.1)}{\leq} \| z^k - z^* \|^2 - \| z^{k+1} - z^* \|^2. \quad (3.4.4)$$

Now we use the identity $z^k = x^k_g + \gamma \nabla g(x^k_g)$ and the Lipschitz continuity of $\nabla g$ to upper bound $\| z^k - z^* \|^2$ by a multiple of $\| x^k_g - x^* \|^2$:

$$\| z^k - z^* \|^2 \leq \left(1 + \frac{\gamma}{\beta_g}\right)^2 \| x^k_g - x^* \|^2. \quad (3.4.5)$$

Rearrange Equation (3.4.4), plug in Equation (3.4.5), and take a square root to get Equation (3.4.6).

\[QED\]

Remark 3.4.1. For all $\lambda \in [0, 1]$, the constant $C(\lambda)$ is minimal when $\gamma = \beta_g$, i.e. $C(\lambda) = (1 - \lambda_k \mu_g \beta_g)^{1/2}$. Furthermore, for any choice of $\gamma$, we have the bound $C(1) \leq C(\lambda)$. In particular, for $g = (1/2)\| \cdot \|^2$, the PRS algorithm converges in one step ($C(1) = 0$). Thus, this rate is tight.

The following theorem deduces linear convergence of relaxed PRS whenever $f$ carries both regularity properties. Note that linear convergence of the PRS algorithm ($\lambda_k \equiv 1$) does not follow.

Theorem 3.4.2 (Linear convergence with regularity of $f$). Let $z^*$ be a fixed point of $T_{PRS}$, let $x^* = \text{prox}_{\gamma g}(z^*)$, and suppose that $\mu_f \beta_f > 0$. For all $\lambda \in [0, 1]$, let

$$C(\lambda) := \left(1 - (\lambda/2) \min \left\{ 4\gamma \mu_f / (1 + \gamma / \beta_f)^2, (1 - \lambda) \right\} \right)^{1/2}.$$

Then for all $k \geq 0$,

$$\| z^{k+1} - z^* \| \leq C(\lambda_k) \| z^k - z^* \|. \quad (3.4.6)$$

Proof. Theorem 3.2.1 bounds the distance of $x^k_f$ to the minimizer (where we substitute $z^{k+1} - z^k = 2\lambda_k (x^k_f - x^k_g)$)

$$4\gamma \lambda_k \mu_f \| x^k_f - x^* \|^2 + 4\lambda_k (1 - \lambda_k) \| x^k_f - x^k_g \|^2 \overset{(3.2.1)}{\leq} \| z^k - z^* \|^2 - \| z^{k+1} - z^* \|^2. \quad (3.4.7)$$
Recall the identities:

\[ z^k = x^k_g + \gamma \nabla g(x^k_g) = x^k_f - \gamma \nabla f(x^k_f) + 2(x^k_g - x^k_f) \quad \text{and} \quad z^* = x^* - \gamma \nabla f(x^*). \]

Therefore, by the convexity of \( \| \cdot \|^2 \), we can bound the distance of \( z^k \) to the fixed point \( z^* \)

\[ \| z^k - z^* \|^2 \leq 2 \left( \left( 1 + \frac{\gamma}{\beta_f} \right)^2 \| x^k_f - x^* \|^2 + 4 \| x^k_g - x^k_f \|^2 \right). \]  (3.4.8)

Equations (3.4.7) and (3.4.8) produce the contraction:

\[
C' \| z^k - z^* \|^2 + \| z^{k+1} - z^* \|^2 \\
\leq 4 \gamma \lambda_k \mu_f \| x^k_f - x^* \|^2 + 4 \lambda_k \| x^k_f - x^k_g \|^2 + \| z^{k+1} - z^* \|^2 \\
\leq 2 \| z^k - z^* \|^2
\]

where \( C' = (\lambda_k / 2) \min \{ 4 \gamma \mu_f / (1 + \gamma / \beta_f)^2, (1 - \lambda_k) \} \). Therefore, Equation (3.4.6) holds.

\[
3.4.2 \quad \text{Complementary regularity of } f \text{ and } g
\]

In this subsection, we assume that \( f \) and \( g \) share the regularity. In symbols: \( \mu f \beta_g + \mu g \beta_f > 0 \). In this case, linear convergence is expected. To the best of our knowledge, the next result is new.

**Theorem 3.4.3** (Linear convergence: mixed case). Let \( z^* \) be a fixed point of \( T_{\text{PRS}} \), let \( x^* = \text{prox}_{\gamma g} (z^*) \), and suppose that \( \nabla g \), (respectively \( \nabla f \)), is \((1/\beta)\)-Lipschitz and \( f \), (respectively \( g \)), is \( \mu \)-strongly convex. For all \( \lambda \in [0, 1] \), let \( C(\lambda) := (1 - (4 \lambda / 3) \min \{ \gamma \mu, \beta / \gamma, (1 - \lambda) \})^{1/2} \). Then for all \( k \geq 0 \),

\[
\| z^{k+1} - z^* \| \leq C(\lambda_k) \| z^k - z^* \|. 
\]

**Proof.** First assume that \( \mu f \beta_g > 0 \). Theorem 3.2.1 bounds the distance of \( x^k_f \) to the minimizer and the distance of \( \nabla g(x^k_g) \) to the optimal gradient (where we substitute \( z^{k+1} - \)
\[ z^k = 2\lambda_k (x^k_f - x^k_g) \]:

\[
4\gamma\lambda_k \mu \|x^k_f - x^*\|^2 + 4\gamma\lambda_k \beta \|\nabla g(x^k_g) - \nabla g(x^*)\|^2 + 4\lambda_k (1 - \lambda_k) \|x^k_f - x^k_g\|^2 \leq \|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2. \tag{3.4.9}
\]

Recall the identities:

\[ z^k = x^k_g + \gamma \nabla g(x^k_g) = x^k_f + \gamma \nabla g(x^k_g) + (x^k_g - x^k_f) \quad \text{and} \quad z^* = x^* + \gamma \nabla g(x^*). \]

Thus, from the convexity of \(\|\cdot\|^2\),

\[
\|z^k - z^*\|^2 \leq 3 (\|x^k_f - x^*\|^2 + \|\gamma \nabla g(x^k_g) - \gamma \nabla g(x^*)\|^2 + \|x^k_g - x^k_f\|^2). \tag{3.4.10}
\]

We use Equation (3.4.10) to bound the distance of \(z^k\) to the fixed point \(z^*\) by the left hand side of Equation (3.4.9):

\[
C'\|z^k - z^*\|^2 \\
\leq 4\gamma\lambda_k \mu \|x^k_f - x^*\|^2 + 4\lambda_k (1 - \lambda_k) \|x^k_f - x^k_g\|^2
\]

where \(C' = (4\lambda_k/3) \min\{\gamma \mu, \beta/\gamma, (1 - \lambda_k)\}\). Therefore, we reach the contraction:

\[
\|z^{k+1} - z^*\| \leq (1 - (4\lambda_k/3) \min\{\gamma \mu, \beta/\gamma, (1 - \lambda_k)\})^{1/2} \|z^k - z^*\|^2.
\]

If \(\mu g \beta_f > 0\), then the proof is nearly identical, but relies on the identity:

\[ z^k = x^k_g + \gamma \tilde{\nabla} g(x^k_g) = x^k_g - \gamma \nabla f(x^k_f) + (x^k_g - x^k_f). \]

\[ \square \]

3.5 Feasibility Problems with regularity

In this section we consider the feasibility problem:

Given two closed convex subsets \(C_f\) and \(C_g\) of \(H\) such that \(C_f \cap C_g \neq \emptyset\), find a point \(x \in C_f \cap C_g\).
Throughout this section we assume that \( \{C_f, C_g\} \) is \textit{boundedly linearly regular}:

**Definition 3.5.1** (Bounded linear regularity). Suppose that \( C_1, \ldots, C_m \) are closed convex subsets of \( \mathcal{H} \) with nonempty intersection. We say that \( \{C_1, \ldots, C_m\} \) is boundedly linearly regular if the following holds: for all \( \rho > 0 \), there exists \( \mu_\rho > 0 \) such that for all \( x \in B(0, \rho) \), (the open ball centered at the origin with radius \( \rho \)), we have

\[
d_{C_1 \cap \cdots \cap C_m}(x) \leq \mu_\rho \max\{d_{C_f}(x), \ldots, d_{C_m}(x)\}
\]

where for any subset \( C \subseteq \mathcal{H} \), the distance function \( d_C(x) \) is defined in Equation (3.1.4). Evidently, if \( B(0, \rho) \setminus (C_1 \cap \cdots \cap C_m) \neq \emptyset \), then \( \mu_\rho \geq 1 \).

We say that \( \{C_1, \ldots, C_m\} \) is linearly regular if it is boundedly linearly regular and \( \mu_\rho \) does not depend on \( \rho \), i.e. \( \mu_\rho = \mu_\infty < \infty \).

Intuitively, (bounded) linear regularity is the following implication:

\[
(\text{close to all of the sets}) \implies (\text{close to the intersection})
\]

This property will be key to deducing linear convergence of an application of the relaxed PRS algorithm. See Section 2.9.1 of Chapter 2 for the feasibility problem when no regularity is assumed.

There are several ways to model the feasibility problem, e.g. with \( f \) and \( g \) given by indicator functions, distance functions, or squared distance functions. In this section, we will model the feasibility problem using squared distance functions:

\[
f(x) := d_{C_f}^2(x) \quad \text{and} \quad g(x) := d_{C_g}^2(x).
\]

We briefly summarize some properties of squared distance functions.

**Proposition 3.5.1** (Properties of distance functions). Let \( C \) be a nonempty closed convex subset of \( \mathcal{H} \). Then the following properties hold:

1. The function \( d_C \) is 1-Lipschitz.
2. The function $d_C^2$ is differentiable, and $\nabla d_C^2 = 2(I_H - P_C)$. In addition, $\nabla d_C^2$ is 2-Lipschitz.

3. The proximal identity holds: for all $\gamma > 0$,

$$\text{prox}_{\gamma d_C^2} = \frac{1}{2\gamma + 1} I_H + \frac{2\gamma}{2\gamma + 1} P_C.$$  

Proof. For a proof see [11, Corollary 12.30].

In this section, we will vary the implicit stepsize parameter in every iteration. In addition $f$ and $g$ will have separate implicit stepsize parameters. Thus, we augment the $T_{PRS}$ notation as follows: for all $\gamma_f, \gamma_g > 0$,

$$T_{PRS}^{\gamma_f, \gamma_g} := \text{refl}_{\gamma_f} \circ \text{refl}_{\gamma_g}.$$  

The following propositions study the behavior of $T_{PRS}^{\gamma_f, \gamma_g}$ as the positive implicit stepsize parameters $\gamma_f$ and $\gamma_g$ vary.

**Lemma 3.5.1** (Non expansiveness of PRS operator). The operator $T_{PRS}^{\gamma_f, \gamma_g}$ is nonexpansive.

Proof. This is an immediate consequence of the nonexpansiveness of the reflection mapping (See Part 3 of Proposition 3.1.1).

The following lemma will be useful for determining the fixed point set of $T_{PRS}^{\gamma_f, \gamma_g}$.

**Lemma 3.5.2** (Minimizers of weighted squared distance). Let $\rho_1, \rho_2 > 0$, and suppose that $C_f \cap C_g \neq \emptyset$. Then the set of minimizers of $\rho_1 d_{C_f}^2 + \rho_2 d_{C_g}^2$ is $C_f \cap C_g$.

Proof. The minimal value is attained whenever $x \in C_f \cap C_g$; otherwise, the sum is nonzero.

We will now compute the fixed points of $T_{PRS}^{\gamma_f, \gamma_g}$.

**Proposition 3.5.2** (Fixed points of PRS operator). The set of fixed points of $T_{PRS}^{\gamma_f, \gamma_g}$ is $C_f \cap C_g$. 

101
Proof. Let \( f' = \gamma_f f \) and let \( g' = \gamma_g g \). Then Lemma 3.1.3 combined with Lemma 3.5.2 show that the set of fixed points of \( T_{PRS}^{\gamma_f, \gamma_g} = \text{refl}_{f'} \circ \text{refl}_{g'} \) is
\[
\{ x + \gamma \nabla g'(x) \mid x \in C_f \cap C_g, \nabla g'(x) = -\nabla f'(x) \}.
\]
However, \( \nabla g'(x) = 2\gamma_g (x - P_{C_g}(x)) = 0 \) for all \( x \in C_f \cap C_g \), and so the identity holds. \( \square \)

Given \( z^0 \in H \), sequences of implicit stepsizes parameters, \( (\gamma_{f,j})_{j \geq 0}, (\gamma_{g,j})_{j \geq 0} \), and relaxation parameters, \( (\lambda_j)_{j \geq 0} \), we consider the iteration: for all \( k \geq 0 \), let
\[
\begin{align*}
x^k_f & = \text{prox}_{\gamma_{g,k} d^2_{C_g}}(z^k); \\
x^k_g & = \text{prox}_{\gamma_{f,k} d^2_{C_f}}(2x^k_f - z^k); \\
z^{k+1} & = z^k + 2\lambda_k(x^k_f - x^k_g).
\end{align*}
\] (3.5.1)

If \( (\gamma_{f,j})_{j \geq 0}, (\gamma_{g,j})_{j \geq 0} \subseteq (0, 1/2] \) and \( \lambda_k \equiv 1 \), then the iteration in Equation (3.5.1) is the underrelaxed MAP (see [9] for the parallel product space version and see [4] for the nonconvex case). In particular, Corollary 3.5.1 (below) shows that when all implicit stepsizes parameters are equal to 1/2 and all relaxation parameters are 1, Equation (3.5.1) reduces to the MAP algorithm, where \( x^k_f = P_{C_f} z^k = 2x^k_g - z^k \), and \( z^{k+1} = P_{C_f} P_{C_g} z^k \). This was already noticed in [87, Proposition 2.5] for the fixed \( \gamma \) case.

Now, we will show that the sequence generated by Equation (3.5.1) is bounded.

**Proposition 3.5.3** (Boundedness). Suppose that \( (z^k) \) is generated by the iteration in Equation (3.5.1). If \( (\lambda_k)_{k \geq 0} \subseteq (0, 1] \), then \( (\|z^j - x^j\|^2)^{j \geq 0} \) is monotonically decreasing for any \( x \in C_f \cap C_g \).

*Proof.* Because the set of fixed points of \( T_{PRS}^{\gamma_f, \gamma_g} \) does not depend on \( \gamma_f \) and \( \gamma_g \), the claim follows directly from the Fejér-type inequality in Equation (3.1.12). \( \square \)

**Proposition 3.5.4** (Upper fundamental inequality for feasibility problem). Suppose that \( z \in H \) and \( z^+ = (T_{PRS}^{\gamma_f, \gamma_g})_\lambda(z) \). Then for all \( x^* \in C_f \cap C_g \),
\[
8\lambda(\gamma_f d^2_{C_f}(x_f) + \gamma_g d^2_{C_g}(x_g)) \leq \|z - x^*\|^2 - \|z^+ - x^*\|^2 + \left(1 - \frac{1}{\lambda}\right)\|z^+ - z\|^2. \quad (3.5.2)
\]
Proof. This follows directly from the upper fundamental inequality in Proposition 3.1.3 (with $\mu_f = \mu_g = 0$, and $\gamma = 1$), applied to the functions $f' = \gamma_1 f$ and $g' = \gamma_2 g$. Indeed, the gradients $\gamma_f \nabla d^2_{C_f}$ and $\gamma_g \nabla d^2_{C_g}$ are $2\gamma_f$ and $2\gamma_g$-Lipschitz ($\beta_f = 1/(2\gamma_f)$ and $\beta_g = 1/(2\gamma_g)$). Furthermore, if $S_{g'}$ and $S_{f'}$ are defined as in Equation (3.1.19), then

$$S_{g'}(x_{g'}, x^*) = \frac{1}{4\gamma_g} \| \gamma_g \nabla d^2_{C_g}(x_g) - \gamma_g \nabla d^2_{C_g}(x^*) \|^2 = \frac{\gamma_g}{4} \| (x_g - P_{C_g}(x_g)) \|^2 = \gamma_g d^2_{C_g}(x_g),$$

and by the same argument, $S_{f'}(x_{f'}, x^*) = \gamma_f d^2_{C_f}(x_f)$. To summarize, we have

$$\gamma_f d^2_{C_f}(x_f) + \gamma_g d^2_{C_g}(x_g) + S_{f'}(x_{f'}, x^*) + S_{g'}(x_{g'}, x^*) = 2\gamma_f d^2_{C_f}(x_f) + 2\gamma_g d^2_{C_g}(x_g). \tag{3.5.4}$$

Therefore, the inequality follows because $d^2_{C_g}(x^*) = d^2_{C_f}(x^*) = 0$. \hfill \Box

We are now ready to prove the linear convergence of Algorithm (3.5.1) whenever $\{C_f, C_g\}$ is (boundedly) linearly regular. The proof is a consequence of the upper inequality in Proposition 3.5.4.

**Theorem 3.5.1 (Linear convergence: Feasibility for two sets).** Suppose that $(z^j)_{j\geq 0}$ is generated by the iteration in Equation (3.5.1), and that $C_f$ and $C_g$ are (boundedly) linearly regular. Let $\rho > 0$ and $\mu > 0$ be such that $(z^j)_{j\geq 0} \subseteq B(0, \rho)$ and the inequality

$$d_{C_f \cap C_g}(x) \leq \mu \max \{d_{C_f}(x), d_{C_g}(x)\}$$

holds for all $x \in B(0, \rho)$. Then $(z^j)_{j\geq 0}$ satisfies the following relation: for all $k \geq 0$,

$$d_{C_f \cap C_g}(z^{k+1}) \leq C(\gamma_{f,k}, \gamma_{g,k}, \lambda_k, \mu) \times d_{C_f \cap C_g}(z^k) \tag{3.5.5}$$

where

$$C(\gamma_{f,k}, \gamma_{g,k}, \lambda_k, \mu) := \left(1 - \frac{4\lambda_k \min \{\gamma_{g,k}/(2\gamma_g + 1)^2, \gamma_{f,k}/(2\gamma_f + 1)^2\}}{\mu \max \{16\gamma^2_{g,k}/(2\gamma_g + 1)^2, 1\}}\right)^{1/2}.$$

In particular, if $\overline{C} = \sup_{j \geq 0} C(\gamma_{f,j}, \gamma_{g,j}, \lambda_j, \mu) < 1$, then $(z^j)_{j\geq 0}$ converges linearly to a point in $x \in C_f \cap C_g$ with rate $\overline{C}$, and

$$\|z^k - x\| \leq 2d_{C_f \cap C_g}(z^0) \prod_{i=0}^{k} C(\gamma_{f,i}, \gamma_{g,i}, \lambda_i, \mu). \tag{3.5.6}$$
Proof. For simplicity, throughout the proof we will drop the iteration index $k$ and denote $z^+ := z^{k+1}$ and $z := z^k$, etc. Now recall the identities:

$$x_g = \frac{1}{2\gamma_g + 1} z + \frac{2\gamma_g}{2\gamma_g + 1} P_{C_g}(z);$$

$$x_f = \frac{1}{2\gamma_f + 1} \text{refl}_{\gamma_g}(z) + \frac{2\gamma_f}{2\gamma_f + 1} P_{C_f}(\text{refl}_{\gamma_g}(z)).$$

Thus, $x_g$ is a point on the line segment connecting $P_{C_g}(z)$ and $z$, and $x_f$ is a point on the line segment connecting $\text{refl}_{\gamma_g}(z)$ and $P_{C_f}(\text{refl}_{\gamma_g}(z))$. Hence, we have the projection identities: $P_{C_g} z = P_{C_g} x_g$ and $P_{C_f}(\text{refl}_{\gamma_g}(z)) = P_{C_f} x_f$. We can also compute the distances to $C_f$ and $C_g$:

$$d^2_{C_g}(x_g) = \frac{1}{(2\gamma_g + 1)^2} d^2_{C_g}(z)$$

and

$$d^2_{C_f}(x_f) = \frac{1}{(2\gamma_f + 1)^2} d^2_{C_f}(\text{refl}_{\gamma_g}(z)).$$

(3.5.7)

We will now bound $d^2_{C_f}(z)$. Because $x_g$ is a point on the line segment connecting $z$ and $P_{C_g}(z)$, Equation (3.5.7) shows that that $\|z - x_g\| = (2\gamma_g/(2\gamma_g + 1)) d_{C_g}(z)$. Thus, if $c_1 := c_1(\gamma_g) = 4\gamma_g/(2\gamma_g + 1)$, we have

$$\|z - \text{refl}_{\gamma_g}(z)\| = 2\|z - x_g\| = c_1 d_{C_g}(z).$$

(3.5.8)

Therefore, because $d_{C_f}$ is 1-Lipschitz and by the convexity of $(\cdot)^2$,

$$d^2_{C_f}(z) \leq (\|z - \text{refl}_{\gamma_g}(z)\| + d_{C_f}(\text{refl}_{\gamma_g}(z)))^2$$

$$= (c_1 d_{C_g}(z) + d_{C_f}(\text{refl}_{\gamma_g}(z)))^2$$

$$\leq 2 \max\{c_1^2, 1\} (d^2_{C_g}(z) + d^2_{C_f}(\text{refl}_{\gamma_g}(z))).$$

(3.5.9)

Now we will simplify the upper bound in Equation (3.5.2) by using Equation (3.5.7)

$$8\lambda \left( \frac{\gamma_g}{(2\gamma_g + 1)^2} d^2_{C_f}(\text{refl}_{\gamma_g}(z)) + \frac{\gamma_f}{(2\gamma_f + 1)^2} d^2_{C_g}(z) \right)$$

$$+ \|z^+ - x\|^2 + \left( \frac{1}{\lambda} - 1 \right) \|z^+ - z\|^2 \leq \|z - x\|^2.$$

(3.5.10)
Because $1/(2 \max\{c_1^2, 1\}) < 1$, we have

\[
8\lambda \left( \frac{\gamma_g}{(2\gamma_g + 1)^2} d_{C_f}^2(\text{refl}_{\gamma_g}(z)) + \frac{\gamma_f}{(2\gamma_f + 1)^2} d_{C_g}^2(z) \right) \\
\geq 8\lambda \min \left\{ \frac{\gamma_g}{(2\gamma_g + 1)^2}, \frac{\gamma_f}{(2\gamma_f + 1)^2} \right\} \left( d_{C_f}^2(\text{refl}_{\gamma_g}(z)) + d_{C_g}^2(z) \right) \\
\geq \frac{8\lambda \min \{\gamma_g/(2\gamma_g + 1)^2, \gamma_f/(2\gamma_f + 1)^2\}}{2 \max\{c_1^2, 1\}} \max\{d_{C_f}^2(z), d_{C_g}^2(z)\}.
\]  

(3.5.9)

(3.5.10)

(3.5.11)

Now, recall the bounded linear regularity property: for all $x \in B(0, \rho)$,

\[
d_{C_f \cap C_g}(x) \leq \mu_\rho \max\{d_{C_f}(x), d_{C_g}(x)\}.
\]

Thus, for all $x \in C_f \cap C_g$, the lower bound in Equation (3.5.11) shows that (where we use $(1/\lambda - 1) \leq 0$ in Equation (3.5.10))

\[
\frac{4\lambda \min \{\gamma_g/(2\gamma_g + 1)^2, \gamma_f/(2\gamma_f + 1)^2\}}{\mu_\rho^2 \max\{c_1^2, 1\}} d_{C_f \cap C_g}^2(z) + \|z^+ - x\|^2 \\
\leq \|z - x\|^2.
\]

Hence, if we define

\[
C(\gamma_f, \gamma_g, \lambda, \mu_\rho) = \left( 1 - \frac{4\lambda \min \{\gamma_g/(2\gamma_g + 1)^2, \gamma_f/(2\gamma_f + 1)^2\}}{\mu_\rho^2 \max\{c_1^2, 1\}} \right)^{1/2}
\]

and $x = P_{C_f \cap C_g}(z)$, then $d_{C_f \cap C_g}(z) = \|z - x\|$ and $d_{C_f \cap C_g}(z^+) \leq \|z^+ - x\|$. Therefore,

\[
d_{C_f \cap C_g}(z^+) \leq C(\gamma_f, \gamma_g, \lambda, \mu_\rho)d_{C_f \cap C_g}(z).
\]

(3.5.12)

Linear convergence of $(z^j)_{j \geq 0}$ to a point in $C_f \cap C_g$ follows from [11, Theorem 5.12]. The rate follows from Equation (3.5.5).

**Remark 3.5.1.** The recent papers [13, 102] have proved linear convergence of DRS applied to $f = \iota_{C_f}$ and $g = \iota_{C_g}$ under the same bounded linear regularity assumption on the pair $\{C_f, C_g\}$. In [13], the proof uses the FPR to bound the distance of $z^k$ to the fixed point set of $T_{\text{PRS}}$. Note that for any closed convex set $C$, we have the limit: $\text{prox}_{1/\gamma \iota_C}(x) \to P_C(x)$ as $\gamma \to \infty$. Thus, the results of [13] and [102] can be seen as the limiting case of our results, but cannot be recovered from Theorem 3.5.1. Indeed, for any positive $\lambda$ and $\mu$, we have the limit: $C(\gamma', \gamma, \lambda, \mu) \to 1$, as $\gamma, \gamma' \to \infty$. 

105
Remark 3.5.2. The constant $C(\gamma, \gamma', \lambda, \mu)$ has the following form:

$$C(\gamma', \gamma, \lambda, \mu) = \begin{cases} 
(1 - \frac{\lambda(2\gamma+1)^2 \min\{\gamma/(2\gamma+1)^2, \gamma'/((2\gamma'+1)^2)\}}{4\gamma^2 \mu^2})^{1/2}, & \text{if } \gamma \geq \frac{1}{2}; \\
(1 - \frac{4\lambda \min\{\gamma/(2\gamma+1)^2, \gamma'/((2\gamma'+1)^2)\}}{\mu^2})^{1/2}, & \text{otherwise}.
\end{cases}$$

For fixed positive $\gamma, \lambda$ and $\mu$, the function $C(\gamma', \gamma, \lambda, \mu)$ is minimized when $\gamma' = \frac{1}{2}$. Furthermore, it follows that $C(1/2, \gamma, \lambda, \mu)$ is minimized over $\gamma$, at $\gamma = 1/2$. Finally, note that $C(\gamma', \gamma, \lambda, \mu)$ is monotonically decreasing in $\lambda$ and monotonically increasing in $\mu$. Thus, in view of Corollary 3.5.1, we achieve the minimal constant for MAP: $C(1/2, 1/2, 1, \mu) = (1 - 1/(2\mu^2))^{1/2}$.

We can use Theorem 3.5.1 to deduce the linear convergence of MAP and give an explicit rate. In [59, Theorem 3.15], the authors show that $\mu$-linear regularity of a finite collection of sets is equivalent to the linear convergence of the method of cyclic projections applied to these sets. Under the assumption of linear regularity, they derive the rate $(1 - 1/(8\mu^2))^{1/2}$. Corollary 3.5.1 is a special case of one direction of this result but with a better rate. It is not clear whether the rate in [59, Theorem 3.15] can be improved for the general cyclic projections algorithm. The rate we derive in Corollary 3.5.1 appears in [7, Corollary 3.14] under the same assumptions.

Corollary 3.5.1 (Convergence of MAP). Let $(z^j)_{j \geq 0}$ be generated by the iteration in Equation (3.5.1) with $\gamma_{f,k} \equiv \gamma_{g,k} \equiv 1/2$ and $\lambda_k \equiv 1$. Then for all $k \geq 0$, $z^{k+1} = P_{C_f}P_{C_g}z^k$. Thus, MAP is a special case of PRS. Consequently, under the assumptions of Theorem 3.5.1, the iterates of MAP converge linearly to a point in the intersection of $C_f \cap C_g$ with rate $(1 - 1/\mu^2)^{1/2}$.

Proof. Notice that $x^k_g = (1/2)z^k + (1/2)P_{C_g}z^k$ and $\text{refl}_{(1/2)g}(z^k) = P_{C_g}z^k$. Similarly, $x^k_f = (1/2)P_{C_g}(z^k) + (1/2)P_{C_f}P_{C_g}z^k$ and $z^{k+1} = \text{refl}_{(1/2)f}(P_{C_g}z^k) = P_{C_f}P_{C_g}z^k$.

We see that $C(1/2, 1/2, 1, \mu) = (1 - 1/(2\mu^2))^{1/2}$. We can strengthen this rate to $(1 - 1/\mu^2)^{1/2}$ by observing that in Equation (3.5.9) we have $d_{C_f}(z^k) = 0$, and so we can set $c_1 = 0$. The proof then follows the same argument. \qed
Remark 3.5.3. If $C_f$ and $C_g$ are closed subspaces with Friedrichs angle $\cos^{-1}(c_F)$, [10, Corollary 11] shows that $\mu \leq 2/\sqrt{1-c_F}$. Therefore, Corollary 3.5.1 predicts that iterates of MAP converges with rate no less than $((3 + c_F)/4)^{1/2}$. The actual rate for this problem is $c_F^2$. See [6, Section 7] for a comparison between DRS and MAP for two subspaces.

With this interpretation of MAP we can examine the inconsistent case, $C_f \cap C_g = \emptyset$, from a different perspective than the current literature. A part of the following result appeared in [8, Theorem 4.8]. In particular, if $x$ satisfies Equation (3.5.13), then $P_{C_f}x - P_{C_g}x$ is the gap vector of [8, Theorem 4.8].

Corollary 3.5.2 (Convergence of MAP: infeasible case). Let $(z^j)_{j \geq 0}$ be generated by MAP, and suppose that $C_f \cap C_g = \emptyset$. If there exists $x \in H$ such that
\[ x - P_{C_f}x = P_{C_g}x - x, \]  
then $(z^j)_{j \geq 0}$ converges weakly to a point in the following set:
\[ \{ P_{C_f}x \mid x \in H, x - P_{C_f}x = P_{C_g}x - x \} \subseteq C_f, \]  
with FPR rate $\|z^{k+1} - z^k\|^2 = o(1/(k+1))$. Furthermore, if $x$ satisfies Equation (3.5.13), then
\[ \sum_{i=0}^{\infty} \left( \left\| \frac{1}{2}(z^i - P_{C_g}z^i) - (x - P_{C_g}x) \right\|^2 + \left\| \frac{1}{2}(P_{C_g}z^i - P_{C_f}P_{C_g}z^i) - (x - P_{C_f}x) \right\|^2 \right) < \infty. \]  
(3.5.15)

In particular, the vector $P_{C_g}z^k - P_{C_f}P_{C_g}z^k$ strongly converges to the gap vector $P_{C_g}x - P_{C_f}x$, and
\[ \min_{i=0, \ldots, k} \left\{ \left\| (P_{C_g}z^i - P_{C_f}z^i) - (P_{C_g}x - P_{C_f}x) \right\|^2 \right\} = o(1/(k+1)). \]

Proof. In view of Proposition 3.5.1, the condition $x - P_{C_f}x = P_{C_g}x - x$ is equivalent to $x \in \operatorname{zer}(\nabla d_{C_f}^2 + \nabla d_{C_g}^2)$. The mapping $T_{\text{PRS}}^{1/2,1/2} = P_{C_f}P_{C_g}$ is the composition of $(1/2)$-averaged maps, and so it is $\alpha$-averaged for some $\alpha < 1$ [11, Proposition 4.32]. In addition,
Proposition 3.1.2 shows that the FPR satisfies $\|z^{k+1} - z^k\|^2 = o(1/(k + 1))$. The set in Equation (3.5.14) is precisely the set of fixed points of $T_{PRS}$. Therefore, weak convergence follows from [11, Proposition 5.15]. The sum in Equation (3.5.15) is exactly the sum of derivatives $\|\nabla^2 d_{C_g}(x^{k}_g) - \nabla^2 d_{C_g}(x)\|^2 + \|\nabla^2 d_{C_f}(x^{k}_f) - \nabla^2 d_{C_f}(x)\|^2$, and so it is finite by Proposition 3.2.1. Finally, strong convergence of $P_{C_g}z^k - P_{C_f}P_{C_g}z^k$ to the gap vector follows from the identity $x - P_{C_f}x = (1/2)(P_{C_g}x - P_{C_f}x)$. The rate is a consequence of Part 4 of Lemma 3.1.1 and Equation (3.5.15).

Remark 3.5.4. Note that the condition $x - P_{C_f}x = x - P_{C_g}x$ is equivalent to $\|P_{C_g}x - P_{C_f}x\|^2 = \min_{y \in H}(d^2_{C_f}(y) + d^2_{C_g}(y)) = \min_{x_f \in C_f, x_g \in C_g} \|x_g - x_f\|^2$. See [8, Fact 5.1] for conditions that guarantee the infimum is attained in Corollary 3.5.2.

3.5.1 Multiple sets

The concept of (bounded) linear regularity is defined for any finite number of sets. The following theorem shows that (bounded) linear regularity of a collection of sets is equivalent to the (bounded) linear regularity of a certain pair of sets in a product space. For convenience we set

$$D := \{(x, \cdots, x) \mid x \in H \} \subseteq H^m, \quad (3.5.16)$$

and endow $H^m$ with the canonical norm: $\|(x_1, \cdots, x_m)\|^2 = (1/m)\sum_{i=1}^m \|x_i\|^2$. We will use the boldface notation $\mathbf{x} \in H^m$ for an arbitrary vector in $H^m$. Finally, for any $\mathbf{x} \in H^m$, we will write $\mathbf{x}_j$ for the $j$th component of $\mathbf{x}$, which is an element of $H$.

Theorem 3.5.2 ((Bounded) linear regularity in product spaces). Suppose that $C_1, \cdots, C_m$ are closed convex subsets of $H$ with nonempty intersection. Then $\{C_1, \cdots, C_m\}$ is boundedly linearly regular or linearly regular, if, and only if, $\{C_1 \times \cdots \times C_m, D\}$ has the same property in $H^m$ with the canonical norm. In particular, if $\{C_1, \cdots, C_m\}$ is $\mu_{\rho}$-(boundedly) linearly regular on the ball $B(0, \rho)$, then $\{C_1 \times \cdots \times C_m, D\}$ is $\sqrt{(1 + 4m\mu_{\rho}^2)}$-(boundedly) linearly regular.
on the ball $B(0, \rho)$, and
\[
d_{(C_1 \times \cdots \times C_m) \cap D}(x) \leq \sqrt{(1 + 4m\mu^2)} \max\{d_{C_1 \times \cdots \times C_m}(x), d_D(x)\}.
\] (3.5.17)

**Proof.** See [59, Theorem 3.12].

In this section we model the feasibility problem of the $m$ sets $\{C_1, \cdots, C_m\}$ using the following two objective functions on the product space $H^m$:
\[
f(x_1, \cdots, x_m) = \sum_{i=1}^{m} d_{C_i}^2(x_i) \quad \text{and} \quad g(x_1, \cdots, x_n) = d_D^2(x_1, \cdots, x_n).
\]

In the space $H^m$, the proximal operators of $f$ and $g$ have the following form:
\[
\text{prox}_{\gamma f}(x) = \left( \frac{1}{2\gamma + 1} x_i + \frac{2\gamma}{2\gamma + 1} P_{C_j} x_j \right)_{j=1}^{m};
\]
\[
\text{prox}_{\gamma g}(x) = \left( \frac{1}{2\gamma + 1} x_j + \frac{2\gamma}{(2\gamma + 1)m} \sum_{i=1}^{m} x_i \right)_{j=1}^{m}.
\]

We apply the iteration in Equation (3.5.1) with these identities to get the following parallel algorithm: given implicit stepsize parameters $(\gamma_{f,i})_{j \geq 0}$ and $(\gamma_{g,i})_{j \geq 0}$, relaxation parameters $(\lambda_i)_{j \geq 0} \subseteq (0, 1]$, and an initial point $z^0 \in H^m$, for all $k \geq 0$, define
\[
\bar{z}^k = \frac{1}{m} \sum_{i=1}^{m} z_i^k;
\]
\[
\begin{cases}
x_{g,i}^k = (1/(2\gamma_{g,k} + 1))z_i^k + (2\gamma_{g,k}/(2\gamma_{g,k} + 1))z^k; & \text{For } i = 1, \cdots, m \\
x_{f,i}^k = (1/(2\gamma_{f,k} + 1))(2x_{g,i}^k - z_i^k) + (2\gamma_{f,k}/(2\gamma_{f,k} + 1))P_{C_i}(2x_{g,i}^k - z_i^k); & \text{in parallel.}
\end{cases}
\]
\[
z_i^{k+1} = z_i^k + 2\lambda_k(x_{f,i}^k - x_{g,i}^k);
\]
(3.5.18)

Note that the algorithm in Equation (3.5.18) is related to the general algorithm in [5, Section 8.3]. One of the main differences between these two algorithms is that the projection operators are not necessarily evaluated at same point in each iteration ($(2x_{g,i}^k - z^k) \notin D$).

By changing the metric of the underlying space, e.g. to $\|x_1, \cdots, x_m\|^2 = \sum_{i=1}^{m} w_i \|x_i\|^2$
where $w_i > 0$ are arbitrary weights, we can perform a weighted average of all the projections. In addition, we can assign each set $C_i$ a different implicit stepsize parameter at each iteration. For simplicity we do not pursue these extensions here.

The following theorem deduces the linear convergence of the iteration in Equation (3.5.18).

**Theorem 3.5.3** (Linear convergence: Feasibility for multiple sets). Suppose that $(z^j)_{j \geq 0}$ is generated by the iteration in Equation (3.5.18), and suppose that the collection $\{C_1, \ldots, C_m\}$ is (boundedly) linearly regular. Let $\rho > 0$ and $\mu_\rho > 0$ be such that $(z^j)_{j \geq 0} \subseteq B(0, \rho)$ and the inequality

$$d_{C_1 \cap \cdots \cap C_m}(x) \leq \mu_\rho \max\{d_{C_1}(x), \ldots, d_{C_m}(x)\}$$

(3.5.19)

holds for all $x \in B(0, \rho)$. Then $(z^j)_{j \geq 0}$ satisfies the following relation: for all $k \geq 0$,

$$d_{(C_1 \times \cdots \times C_m) \cap D}(z^{k+1}) \leq C(\gamma f, k, \gamma g, k, \lambda_k, \mu_\rho) \times d_{(C_1 \times \cdots \times C_m) \cap D}(z^k)$$

(3.5.20)

where

$$C(\gamma f, k, \gamma g, k, \lambda_k, \mu_\rho) := \left(1 - \frac{4\lambda_k \min\{\gamma g, k/(2\gamma g, k + 1)^2, \gamma f, k/(2\gamma f, k + 1)^2\}}{(1 + 4m\mu_\rho^2) \max\{16\gamma g, k^2/(2\gamma g, k + 1)^2, 1\}}\right)^{1/2}.$$

In particular, if $\overline{C} = \sup_{j \geq 0} C(\gamma f, k, \gamma g, k, \lambda_k, \mu_\rho) < 1$, then $(z^j)_{j \geq 0}$ converges linearly to a point in $(C_1 \times \cdots \times C_m) \cap D$ with rate $\overline{C}$, and

$$\|z^k - z^*\| \leq 2d_{(C_1 \times \cdots \times C_m) \cap D}(z^0) \prod_{i=0}^{k} C(\gamma f, i, \gamma g, i, \lambda_i, \mu_\rho).$$

(3.5.21)

**Proof.** This theorem is a direct corollary of Theorem 3.5.1 except that Theorem 3.5.2 is used to calculate the (bounded) linear regularity constant. \qed

Finally we derive the following analogue of Corollary 3.5.1.

**Corollary 3.5.3** (Convergence of MAP: Multiple sets). Let $(z^j)_{j \geq 0}$ be generated by the iteration in Equation (3.5.18) with $\gamma f, k \equiv \gamma g, k \equiv 1/2$ and $\lambda_k \equiv 1$. Define $x^k := (P_D z^k)_1$. Then for all $k \geq 0$,

$$x^{k+1} = \frac{1}{m} \sum_{i=1}^{m} P_{C_i}(x^k).$$

(3.5.22)
Thus, Averaged MAP is a special case of PRS. Consequently, under the assumptions of Theorem 3.5.3, \( x^k \) converges linearly to a point in the intersection \( C_1 \cap \cdots \cap C_m \) with rate \( (1 - 1/(1 + 4m\mu^2))^{1/2} \).

**Proof.** Equation (3.5.22) follows because \( \text{refl}_{ig} = P_D \) and \( \text{refl}_{if} = P_{C_1 \times \cdots \times C_m} \). In addition, by the nonexpansiveness of \( P_D \) we have

\[
\|x^k - z^*\|^2 = \|(P_Dz^k)_1 - z_1^*\|^2 = \frac{1}{m} \sum_{i=1}^{k} \|(P_Dz^k)_i - z_i^*\|^2 \leq \|z^k - z^*\|^2.
\]

By Corollary 3.5.1 and Theorem 3.5.2, the sequence \((z^j)_{j\geq 0}\) converges linearly with rate \( (1 - 1/(1 + 4m\mu^2))^{1/2} \). Thus, the rate for \((x^k)_{k\geq 0}\) follows from the rate for \((z^j)_{j\geq 0}\). \(\square\)

### 3.6 From relaxed PRS to ADMM

It is well known that ADMM is equivalent to DRS applied to the Lagrange dual of Problem (3.1.2) [63]. Thus, if we let

\[
d_f(w) := f^*(A^*w) \quad \text{and} \quad d_g(w) := g^*(B^*w) - \langle w, b \rangle,
\]

then relaxed ADMM is equivalent to relaxed PRS applied to the following problem:

\[
\min_{w \in \mathcal{G}} d_f(w) + d_g(w). \tag{3.6.1}
\]

We make two assumptions regarding \( d_f \) and \( d_g \).

**Assumption 3.6.1 (Solution existence).** *Functions \( f, g : \mathcal{H} \to (-\infty, \infty] \) satisfy

\[
\text{zer}(\partial d_f + \partial d_g) \neq \emptyset. \tag{3.6.2}
\]

This is a restatement of Assumption 3.1.2, which we have used in our analysis of the primal case.

**Assumption 3.6.2.** *The following differentiation rule holds:

\[
\partial d_f(x) = A^* \circ (\partial f^*) \circ A \quad \text{and} \quad \partial d_g(x) = B^* \circ (\partial g^*) \circ B.
\]
See [11, Theorem 16.37] for conditions that imply this identity. We need this assumption to compute subgradients of \( d_f \) and \( d_g \).

We are going to review how regularity of the primal functions affects the dual. The next proposition shows how the strong convexity and the differentiability of a closed, proper, and convex function transfer to the dual function.

**Proposition 3.6.1** (Strong convexity and differentiability of the conjugate). Suppose that \( f : \mathcal{H} \to (-\infty, \infty] \) is closed, proper, and convex. Then the following implications hold:

1. If \( f \) is \( \mu_f \)-strongly convex, then \( f^* \) is differentiable and \( \nabla f \) is \( (1/\mu_f) \)-Lipschitz.

2. If \( f \) is differentiable and \( \nabla f \) is \( (1/\beta) \)-Lipschitz, then \( f^* \) is \( \beta \)-strongly convex.

**Proof.** See [11, Theorem 18.15].

With Proposition 3.6.1, we can characterize the strong convexity and differentiability of the dual functions in terms of \( A, B \) and \( f \) and \( g \). We first recall that a linear map \( L : \mathcal{G} \to \mathcal{G} \) is \( \alpha \)-strongly monotone if for all \( x \in \mathcal{G} \), the bound \( \langle Lx, x \rangle \geq \alpha \| x \|_G^2 \) holds.

**Proposition 3.6.2** (Strong convexity and differentiability of the dual). The following implications hold:

1. If \( \nabla f \), (respectively \( \nabla g \)), is \( (1/\beta) \)-Lipschitz and \( AA^* \) (respectively \( BB^* \)) is \( \alpha \)-strongly monotone, then \( d_f \) (respectively \( d_g \)) is \( \alpha\beta \)-strongly convex.

2. If \( f \) (respectively \( g \)) is \( \mu \)-strongly convex, then \( d_f \) (respectively \( d_g \)) is differentiable and \( \nabla d_f \) (respectively \( \nabla d_g \)) is \( (\|A\|_G^2/\mu) \) (respectively \( (\|B\|_G^2/\mu) \))-Lipschitz.

The proof of Proposition 3.6.2 is straightforward, so we omit it. We note that \( AA^* \) and \( BB^* \) are always 0-strongly monotone. Thus, we assume that \( AA^* \) and \( BB^* \) are \( \alpha_A \) and \( \alpha_B \)-strongly monotone, respectively, while allowing the cases \( \alpha_A = 0 \) and \( \alpha_B = 0 \). In addition, we use the convention that \( \tilde{\nabla} f \) and \( \tilde{\nabla} g \) are always \( (1/\beta_f) \), and \( (1/\beta_g) \)-Lipschitz,
respectively, by allowing the cases $\beta_f = 0$ and $\beta_g = 0$. We carry the following notation throughout the rest of Section 3.6:

$$\mu_{d_f} = \beta_f \alpha_A \geq 0 \text{ and } \mu_{d_g} = \beta_g \alpha_B \geq 0. \quad (3.6.3)$$

Thus, $d_f$ and $d_g$ are $\mu_{d_f}$ and $\mu_{d_g}$-strongly convex, respectively. Finally, we always assume that $f$ and $g$ are $\mu_f$ and $\mu_g$-strongly convex, respectively, by allowing $\mu_f = 0$ and $\mu_g = 0$. We assume that $\|A\|\|B\| \neq 0$, and denote

$$\beta_{d_f} = \frac{\mu_f}{\|A\|^2} \geq 0 \text{ and } \beta_{d_g} = \frac{\mu_g}{\|B\|^2} \geq 0. \quad (3.6.4)$$

If $\beta_{d_f}$ is strictly positive, then $d_f$ is differentiable and $\nabla d_f$ is $(1/\beta_f)$-Lipschitz. An analogous result holds for $d_g$.

Now we apply Algorithm 4 to the dual problem in Equation (3.6.1). Given $z^0 \in \mathcal{H}$, Lemma 3.1.2 shows that we need to compute the following vectors for all $k \geq 0$:

$$\begin{cases}
  w^k_{d_g} = \text{prox}_{\gamma d_g}(z^k); \\
  w^k_{d_f} = \text{prox}_{\gamma d_f}(2w^k_{d_g} - z^k); \\
  z^{k+1} = z^k + 2\lambda_k (w^k_{d_f} - w^k_{d_g}).
\end{cases} \quad (3.6.5)
$$

In order to apply the relaxed PRS algorithm, we need to compute the proximal operators of the dual functions $d_f$ and $d_g$. The proof of the following lemma is straightforward.

**Lemma 3.6.1** (Proximity operators on the dual). Let $w, v \in \mathcal{G}$. Then the update formulas $w^+ = \text{prox}_{\gamma d_f}(w)$ and $v^+ = \text{prox}_{\gamma d_g}(v)$ are equivalent to the following computations:

$$\begin{cases}
  x^+ = \arg \min_{x \in \mathcal{H}_1} f(x) - \langle w, Ax \rangle + \frac{\gamma}{2} \|Ax\|^2; \\
  w^+ = w - \gamma Ax^+; \\
  y^+ = \arg \min_{y \in \mathcal{H}_2} g(y) - \langle v, By - b \rangle + \frac{\gamma}{2} \|By-b\|^2; \\
  v^+ = v - \gamma (By^+ - b),
\end{cases} \quad (3.6.6)$$

respectively. In addition, the subgradient inclusions hold: $A^* w^+ \in \partial f(x^+)$ and $B^* v^+ \in \partial g(y^+)$. Finally, $w^+$ and $v^+$ are independent of the choice of $x^+$ and $y^+$, respectively, even if they are not unique solutions to the minimization subproblems.
<table>
<thead>
<tr>
<th>Function</th>
<th>Primal subgradient</th>
<th>Dual subgradient</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g$</td>
<td>$\tilde{\nabla} g(y^s) = B^* w^s_{d_g}$</td>
<td>$\tilde{\nabla} d_g(w^s_{d_g}) = By^s - b$</td>
</tr>
<tr>
<td>$f$</td>
<td>$\tilde{\nabla} f(x^s) = A^* w^s_{d_f}$</td>
<td>$\tilde{\nabla} d_f(w^s_{d_f}) = Ax^s$</td>
</tr>
</tbody>
</table>

Table 3.1: Overview of the main subgradient identities used throughout Section 3.6. The letter $s$ denotes a superscript (e.g. $s = k$ or $s = \ast$). See Chapter 2 for a proof.

We can use Lemma 3.6.1 to derive the relaxed form of ADMM in Algorithm 5. Note that this form of ADMM eliminates the “hidden variable” sequence $(z^j)_{j \geq 0}$ in Equation (3.6.5). A detailed proof of Proposition 3.6.3 is contained in Appendix A.1.

**Proposition 3.6.3** (Relaxed ADMM). Let $z^0 \in \mathcal{G}$, and let $(z^j)_{j \geq 0}$ be generated by the relaxed DRS algorithm applied to the dual formulation in Equation (3.6.1). Choose $w_{d_g}^{-1} = z^0$, $x^{-1} = 0$ and $y^{-1} = 0$ and $\lambda_{-1} = 1/2$. Then we have the following identities starting from $k = -1$:

$$y^{k+1} = \arg \min_{y \in \mathcal{H}_2} g(y) - \langle w_{d_g}^k, Ax^k + By - b \rangle$$

$$+ \frac{\gamma}{2} \| Ax^k + By - b + (2\lambda_k - 1)(Ax^k + By^k - b) \|^2;$$

$$w_{d_g}^{k+1} = w_{d_g}^k - \gamma(Ax^k + By^{k+1} - b) - \gamma(2\lambda_k - 1)(Ax^k + By^k - b);$$

$$x^{k+1} = \arg \min_{x \in \mathcal{H}_1} f(x) - \langle w_{d_f}^{k+1}, Ax + By^{k+1} - b \rangle + \frac{\gamma}{2} \| Ax + By^{k+1} - b \|^2;$$

$$w_{d_f}^{k+1} = w_{d_f}^{k+1} - \gamma(Ax^{k+1} + By^{k+1} - b).$$

**Remark 3.6.1.** Proposition 3.6.3 proves that $w_{d_f}^{k+1} = w_{d_g}^{k+1} - \gamma(Ax^{k+1} + By^{k+1} - b)$. Recall that by Equation (3.6.5), $z^{k+1} - z^k = 2\lambda_k(w_{d_f}^k - w_{d_g}^k)$. Therefore, it follows that

$$z^{k+1} - z^k = -2\gamma \lambda_k(Ax^k + By^k - b).$$
3.6.1 Converting dual inequalities to primal inequalities

The ADMM algorithm generates 5 sequences of iterates:

\[(z^j)_{j \geq 0}, (w^j_{df})_{j \geq 0}, \text{ and } (w^j_{dg})_{j \geq 0} \subseteq \mathcal{G} \quad \text{and} \quad (x^j)_{j \geq 0} \subseteq \mathcal{H}_1, (y^j)_{j \geq 0} \subseteq \mathcal{H}_2.\]

In this section we recall some inequalities, which were derived in Section 2.8 of Chapter 2, that relate these sequences to each other through the primal and dual objective functions. The proofs all follow from a straightforward application of the Fenchel-Young inequality [11, Proposition 16.9], the identities in Lemma 3.1.2, and the fundamental inequalities in Proposition 3.1.3 and 3.1.4.

In the following propositions, \(z^*\) will denote a fixed point of \(T_{PRS}\). The point \(w^* := \text{prox}_{\gamma d_g}(z^*)\) is a minimizer of the dual problem in Equation (3.6.1). Finally, we let \(x^*\) and \(y^*\) be defined as in Table 3.1.

**Proposition 3.6.4 (Primal values via dual values).** Suppose that \((z^j)_{j \geq 0}\) is generated by Algorithm 5. Then the following identity holds:

\begin{align*}
4\gamma \lambda_k (f(x_k) + g(y_k) - f(x^*) - g(y^*)) &= -4\gamma \lambda_k (d_f(w^k_{df}) + d_g(w^k_{dg}) - d_f(w^*) - d_g(w^*)) \\
&\quad + \left(2 \left(1 - \frac{1}{2\lambda_k}\right) \|z^k - z^{k+1}\|^2 + 2(z^k - z^{k+1}, z^k - z^{k+1})\right). \quad (3.6.8)
\end{align*}

**Proposition 3.6.5 (ADMM primal upper fundamental inequality).** For all \(k \geq 0\), we have the bound

\begin{align*}
4\gamma \lambda_k (f(x_k) + g(y_k) - f(x^*) - g(y^*)) &\leq \|z^k - (z^* - w^*)\|^2 - \|z^{k+1} - (z^* - w^*)\|^2 + \left(1 - \frac{1}{\lambda_k}\right) \|z^k - z^{k+1}\|^2. \quad (3.6.9)
\end{align*}

**Proposition 3.6.6 (ADMM primal lower fundamental inequality).** For all \(x \in \mathcal{H}_1\) and \(y \in \mathcal{H}_2\), we have the bound:

\begin{align*}
f(x) + g(y) - f(x^*) - g(y^*) \geq \langle Ax + By - b, w^* \rangle. \quad (3.6.10)
\end{align*}
Proof. The lower bound follows from the subgradient inequalities:

\[ f(x) - f(x^*) \geq \langle x - x^*, A^* w^* \rangle; \]
\[ g(y) - g(y^*) \geq \langle y - y^*, B^* w^* \rangle. \]

We can add these inequalities together and use the identity, \( Ax^* + By^* = b \), to get Equation (3.6.10) \( \square \)

### 3.6.2 Converting dual convergence rates to primal convergence rates

In this section, we use the inequalities deduced in Section 3.6.1 and the convergence rates proved in previous sections to derive convergence rates for the primal objective error and strong convergence of various quantities that appear in ADMM. In addition, we translate the results of the previous sections and use Proposition 3.6.2 to state all theorems in terms of purely primal quantities.

We recall the definition of the two auxiliary terms (Equation (3.1.19)):

\[
S_d f (w^k, w^*) = \max \left\{ \frac{\beta_f \alpha_A}{2} \| w^k_d - w^* \|^2, \frac{\mu_f}{2\| A \|^2} \| A x^k - A x^* \|^2 \right\}, \tag{3.6.11}
\]
\[
S_d g (w^k, w^*) = \max \left\{ \frac{\beta_g \alpha_B}{2} \| w^k_d - w^* \|^2, \frac{\mu_g}{2\| B \|^2} \| B y^k - B y^* \|^2 \right\}. \tag{3.6.12}
\]

This form readily follows from Table 3.1.

The following is a direct translation of Theorem 3.2.1 to the current setting. Note that any of the Lipschitz, strong convexity, and strong monotonicity constants may be zero.

**Theorem 3.6.1** (Primal differentiability and strong convexity). Suppose that \((z^j)_{j \geq 0}\) is generated by Algorithm 5. Then

1. **Best iterate convergence:** If \((\lambda_j)_{j \geq 0}\) is bounded away from zero, then

\[
\min_{i=0, \ldots, k} \left\{ S_d f (w^i_d, w^*) + S_d g (w^i_d, w^*) \right\} \leq \frac{\| z^0 - z^* \|^2}{8 \gamma \Delta (k + 1)},
\]

and

\[
\min_{i=0, \ldots, k} \left\{ S_d f (w^i_d, w^*) \right\} = o \left( \frac{1}{k+1} \right) \quad \text{and} \quad S_d g (w^k_{d, \text{best}}, w^*) = o \left( \frac{1}{k+1} \right).
\]

116
2. **Ergodic convergence:** Let \( \overline{w}^k_{df} = (1/\Lambda_k) \sum_{i=0}^{k} w^i_{df} \), let \( \overline{w}^k_{dg} = (1/\Lambda_k) \sum_{i=0}^{k} \lambda_i w^i_{dg} \), let \( \overline{x}^k = (1/\Lambda_k) \sum_{i=0}^{k} x^i \), and let \( \overline{y}^k = (1/\Lambda_k) \sum_{i=0}^{k} \lambda_i y^i \). Then

\[
\max \left\{ \beta_f \alpha_A \| \overline{w}^k_{df} - w^* \|^2, \frac{\mu_f}{\| A \|^2} \| A \overline{x}^k - A x^* \|^2 \right\} + \max \left\{ \beta_g \alpha_B \| \overline{w}^k_{dg} - w^* \|^2 + \frac{\mu_g}{\| B \|^2} \| B \overline{y}^k - B y^* \|^2 \right\} \leq \frac{\| z^0 - z^* \|^2}{4\gamma \Lambda_k}.
\]

3. **General convergence:** If \( \tau = \inf_{j \geq 0} \lambda_j (1 - \lambda_j) > 0 \), then

\[
S_f(w^k_{df}, w^*) + S_g(w^k_{dg}, w^*) \leq \frac{\| z^0 - z^* \|^2}{4\gamma \tau (k + 1)},
\]

and \( S_f(w^k_{df}, w^*) + S_g(w^k_{dg}, w^*) = o(1/\sqrt{k + 1}) \).

The following proposition deduces \( o(1/(k + 1)) \) objective error convergence of standard ADMM whenever \( g \) is strongly convex, and \( \gamma \) is small enough.

**Theorem 3.6.2** (Strong convexity of \( g \)). Suppose that \( g \) is \( \mu_g \)-strongly convex. Let \( \lambda_k \equiv 1/2 \), and let \( \gamma < \kappa \beta = \kappa \mu_g / \| B \|^2 \) (see Theorem 3.3.3). Then for all \( k \geq 1 \), we have the constraint violations convergence rate:

\[
\| A x^k + B y^k - b \|^2 \leq \frac{\beta^2 \| w^0_{dg} - w^* \|^2}{\gamma^2 k^2 (1 + \gamma / \beta)^2 (\beta^2 - \gamma^2 / \kappa^2)} \quad \text{and} \quad \| A x^k + B y^k - b \|^2 = o \left( \frac{1}{k^2} \right).
\]

Moreover, the primal objective errors satisfy

\[
\frac{-\beta \| w^* \| \| w^0_{dg} - w^* \|}{\gamma k (1 + \gamma / \beta) \sqrt{(\beta^2 - \gamma^2 / \kappa^2)}} \leq f(x^k) + g(y^k) - f(x^*) - g(y^*) \leq \frac{\beta \| w^0_{dg} - w^* \| (\| z^0 - z^* \|^2 + \| w^* \|^2)}{\gamma k (1 + \gamma / \beta) \sqrt{(\beta^2 - \gamma^2 / \kappa^2)}},
\]

and

\[
| f(x^k) + g(y^k) - f(x^*) - g(y^*) | = o \left( \frac{1}{k} \right).
\]
Proof. The constraint violations rate follows from the identity $z^{k+1} - z^k = -\gamma (Ax^k + By^k - b)$ (Equation (3.6.7)) and the FPR convergence rate in Theorem 3.3.4.

The lower bound follows from the lower fundamental inequality in Proposition 3.6.6 and the FPR convergence rate in Theorem 3.3.4:

$$f(x^k) + g(y^k) - f(x^*) - g(y^*) \geq \langle Ax^k + By^k - b, w^* \rangle$$

$$\geq -\|z^k - z^{k+1}\| \|w^*\|$$

$$\geq -\beta \|w^*\| \|w^0 - w^*\| \gamma k (1 + \gamma/\beta) \sqrt{(\beta^2 - \gamma^2/\kappa^2)}.$$  

Part 1 of Proposition 3.1.2 bounds the norm: $\|z^{k+1} - (z^* - w^*)\| \leq \|z^{k+1} - z^*\| + \|w^*\| \leq \|z^0 - z^*\| + \|w^*\|$. Therefore, the upper bound follows from the upper fundamental inequality in Proposition 3.6.5 and the FPR convergence rate in Theorem 3.3.4:

$$f(x^k) + g(y^k) - f(x^*) - g(y^*) \leq \frac{1}{2\gamma} (\|z^k - (z^* - w^*)\|^2 - \|z^{k+1} - (z^* - w^*)\|^2 - \|z^k - z^{k+1}\|^2)$$

$$\leq \frac{1}{\gamma} \langle z^{k+1} - (z^* - w^*), z^k - z^{k+1} \rangle$$

$$\leq \frac{\beta \|w^0 - w^*\| \|z^0 - z^*\| + \|w^*\| \|z^0 - z^*\| + \|w^*\|} {\gamma k (1 + \gamma/\beta) \sqrt{(\beta^2 - \gamma^2/\kappa^2)}}.$$  

The little $o$-rate follows because the objective error is upper and lower bounded by a multiple of the square root of the FPR, which has convergence rate $o(1/k)$ by Theorem 3.3.4.

It would be nice to prove a convergence rate for the “best iterate” of the sequence of primal objective errors in the style of Theorem 3.3.2. Unfortunately the fundamental inequalities we developed in Section 3.6.1 do not immediately imply such a rate.

Now we shift our focus to linear convergence. The following proposition is a translation of Proposition 3.4.1 to the current setting.
Proposition 3.6.7 (Consequences of linear convergence of ADMM). Let \((C_j)_{j \geq 0} \subseteq [0, 1]\) be a positive scalar sequence, and suppose that for all \(k \geq 0\),
\[
\|z^{k+1} - z^*\| \leq C_k \|z^k - z^*\|. \tag{3.6.13}
\]

Fix \(k \geq 1\). Then
\[
\|w_d^k - w^*\|^2 + \gamma^2 \|By^k - By^*\|^2 \leq \|z^0 - z^*\|^2 \prod_{i=0}^{k-1} C_i^2;
\]
\[
\|w_f^k - w^*\|^2 + \gamma^2 \|Ax^k - Ax^*\|^2 \leq \|z^0 - z^*\|^2 \prod_{i=0}^{k-1} C_i^2.
\]

If \(\lambda < 1\), then the FPR rate holds:
\[
\|(TP_{PR})\lambda z^k - z^k\| \leq \sqrt{\frac{\lambda}{1 - \lambda}} \|z^0 - z^*\| \prod_{i=0}^{k-1} C_i.
\]

Consequently, the following convergence rates for constraint violations and objective errors hold:
\[
\|Ax^k + By^k - b\|^2 \leq \frac{\|z^0 - z^*\|^2}{\gamma^2} \prod_{i=0}^{k-1} C_i^2,
\]
and
\[
-\frac{\|z^0 - z^*\||w^*|}{\gamma} \prod_{i=0}^{k-1} C_i \\
\leq f(x^k) + g(y^k) - f(x^*) - g(y^*) \\
\leq \left(\frac{\|z^0 - z^*\| + \|w^*\|}{\gamma} \right) \|z^0 - z^*\| \prod_{i=0}^{k-1} C_i.
\]

Proof. The convergence rates for the dual variables, primal variables, and FPR follow from Proposition 3.4.1 and the identities in Table 3.1.

Now fix \(k \geq 1\), and let \(z_\lambda = (TP_{PR})\lambda z^k\) for all \(\lambda \in [0, 1]\). The convergence rate for the constraint violation follows from the identity \(z_\lambda - z^k = -2\gamma\lambda(Ax^k + By^k - b)\) (Equation (3.6.7)) and the FPR convergence rate:
\[
\|Ax^k + By^k - b\|^2 = \inf_{\lambda \in [0, 1]} \frac{\|z_\lambda - z^k\|^2}{4\gamma^2\lambda^2} = \inf_{\lambda \in [0, 1]} \frac{\|z^0 - z^*\|^2}{4\gamma^2\lambda(1 - \lambda)} \prod_{i=0}^{k-1} C_i^2 = \frac{\|z^0 - z^*\|^2}{\gamma^2} \prod_{i=0}^{k-1} C_i^2.
\]

119
The lower bound on the objective error follows from the fundamental lower inequality in Proposition 3.6.6 and the constraint violations rate:

\[
\begin{align*}
  f(x^k) + g(y^k) - f(x^*) - g(y^*) & \geq \langle Ax^k + By^k - b, w^* \rangle \\
  & \geq -\|Ax^k + By^k - b\|\|w^*\| \\
  & \geq -\|z^0 - z^*\|\|w^*\| \prod_{i=0}^{k-1} C_i
\end{align*}
\]

The upper bound on the objective error follows from Proposition 3.6.5, the FPR rate, the bound \(\|z_\lambda - z^*\|^2 \leq \|z^k - z^*\|\) (Equation (3.1.12)), the monotonicity of the sequence \(\|z^j - z^*\|_{j \geq 0}\) (Part 1 of Proposition 3.1.2), and the following inequalities:

\[
\begin{align*}
  f(x^k) + g(y^k) - f(x^*) - g(y^*) & \leq \inf_{\lambda \in [0, 1]} \frac{1}{4\gamma\lambda} \left( \|z^k - (z^* - w^*)\|^2 - \|z_\lambda - (z^* - w^*)\|^2 \\
  & \quad + \left(1 - \frac{1}{\lambda}\right) \|z^k - z_\lambda\|^2 \right) \\
  & \leq \frac{1}{4\gamma\lambda} \left( 2\langle z_\lambda - (z^* - w^*), z^k - z_\lambda \rangle + 2 \left(1 - \frac{1}{2\lambda}\right) \|z^k - z_\lambda\|^2 \right) \\
  & \leq \frac{\|z_1/2 - z^*\| + \|w^*\|}{\gamma} \|z_1/2 - z^k\| \\
  & \leq \frac{\|z_0 - z^*\| + \|w^*\|}{\gamma} \|z^0 - z^*\| \prod_{i=0}^{k-1} C_i.
\end{align*}
\]

The following proposition is a direct translation of the main results of Section 3.4 to the current setting.

**Theorem 3.6.3** (Linear convergence of Relaxed ADMM). The following are true:

1. If \(\mu \beta \alpha_B > 0\), then \((z^j)_{j \geq 0}\) converges linearly and

\[
\|z^{k+1} - z^*\|^2 \leq \left(1 - \frac{4\gamma\lambda_k \beta \alpha_B}{(1 + \gamma\|B\|^2/\mu_g)^2}\right)^{1/2} \|z^k - z^*\|^2.
\]
2. If $\mu f \beta f \alpha_A > 0$, then $(z^j)_{j \geq 0}$ converges linearly and

$$
\|z^{k+1} - z^*\|^2 \\
\leq \left( 1 - \frac{\lambda_k \min \left\{ \frac{4 \gamma f \beta f \alpha_A}{(1 + \|A\|^2 / \mu_f)^2}, (1 - \lambda_k) \right\}}{2} \right)^{1/2} \|z^k - z^*\|^2.
$$

3. If $\mu f \beta g \alpha_B > 0$, then $(z^j)_{j \geq 0}$ converges linearly and

$$
\|z^{k+1} - z^*\|^2 \\
\leq \left( 1 - \frac{4 \lambda_k \min \left\{ \frac{\gamma f \beta g \alpha_B}{(1 - \lambda_k)} \right\}}{3} \right)^{1/2} \|z^k - z^*\|^2.
$$

4. If $\mu g f \beta f \alpha_A > 0$, then $(z^j)_{j \geq 0}$ converges linearly and

$$
\|z^{k+1} - z^*\|^2 \\
\leq \left( 1 - \frac{4 \lambda_k \min \left\{ \frac{\gamma f \beta f \alpha_A}{(1 - \lambda_k)} \right\}}{3} \right)^{1/2} \|z^k - z^*\|^2.
$$

We can apply Proposition 3.6.7 to any of the scenarios that appear in Theorem 3.6.3 and deduce the rate of linear convergence of the objective error and constraint violations. We leave this application to the reader.

Linear convergence of ADMM has been deduced in a variety of scenarios. In [58], the authors prove the linear convergence (in finite dimensions) of a generalized form of ADMM, which allows the possibility of adding proximal terms to the alternating minimization steps that appear in Algorithm 5. The four scenarios that appear in [58, Table 1.1]) have some overlap with our results. In the standard version of ADMM, (with no relaxation or extra proximal terms), scenarios 1 and 2 in [58, Table 1.1] are the finite-dimensional analogues of Part 1 of Theorem 3.4. Scenarios 3 and 4 in [58, Table 1.1] are not covered by our analysis because they require that we treat the structure of $A$ and $B$ more carefully than we have in this section. Finally, we note that Parts 2, 3, and 4 of Theorem 3.6.3 are not discussed in [58].

### 3.7 Examples

In this section, we apply DRS and ADMM to concrete problems and explicitly bound the associated objective errors and FPR with the convergence rates that we derived in the
previous sections.

3.7.1  Feasibility problems

Suppose that $C_f$ and $C_g$ are closed convex subsets of $\mathcal{H}$ with nonempty intersection. The goal of the feasibility problem is the find a point in the intersection of $C_f$ and $C_g$. In this section, we present a comparison between MAP and the relaxed PRS algorithm.

3.7.1.1  Linear convergence

Section 3.5 shows that relaxed PRS applied to $f = d^2_{C_f}$ and $g = d^2_{C_g}$ converges linearly whenever $C_f$ and $C_g$ have a sufficiently nice intersection. In addition, [13] and [102] have recently shown that one can achieve linear convergence under the same regularity assumptions on $C_f \cap C_g$ when $f = \chi_{C_f}$ and $g = \chi_{C_g}$. We refer to [13, Fact 5.8] for an extensive list of conditions that guarantee (bounded) linear regularity of $\{C_1, C_2\}$. For the readers convenience, we list a few important examples:

1. **Subspaces:** If $C^\perp_f + C^\perp_g$ is closed, then $\{C_f, C_g\}$ is linearly regular.

2. **Polyhedra:** If $C_f \cap C_g \neq \emptyset$, then $\{C_f, C_g\}$ is linearly regular.

3. **Standard constraint qualification:** If the relative interiors of $C_f$ and $C_g$ intersect, then $\{C_f, C_g\}$ is boundedly linearly regular.

3.7.1.2  General convergence

In general, we cannot expect linear convergence of relaxed PRS algorithm for the feasibility problem. Indeed, Theorem 2.6.2 of Chapter 2 constructs a DRS iteration that converges in norm but does so arbitrarily slowly. A similar result holds for MAP [12]. Thus, in Chapter 2, we focused on other measures of convergence, namely FPR and objective error rate. The following discussion will utilize the results of Chapter 2 to compare the relaxed PRS and MAP algorithms in the absence of regularity.
Let $\chi_{C_f}$ and $\chi_{C_g}$ be the indicator functions of $C_f$ and $C_g$. Then $x \in C_f \cap C_g$, if, and only if, $\chi_{C_f}(x) + \chi_{C_g}(x) = 0$, and the sum is infinite otherwise. Thus, a point is in the intersection of $C_f$ and $C_g$ if, and only if, it is the minimizer of the following problem:

$$\min_{x \in \mathcal{H}} \chi_{C_f}(x) + \chi_{C_g}(x).$$

(3.7.1)

The relaxed PRS algorithm applied to $f = \chi_{C_f}$ and $g = \chi_{C_g}$ has the following form: given an initial point $z^0 \in \mathcal{H}$, for all $k \geq 0$, define

$$\begin{aligned}
x^k_g &= P_{C_g}(z^k); \\
x^k_f &= P_{C_f}(2x^k_g - z^k); \\
z^{k+1} &= z^k + 2\lambda_k(x^k_f - x^k_g).
\end{aligned}$$

(3.7.2)

In general, the functions $f$ and $g$ are neither differentiable nor strongly convex. Furthermore, they only take on the values 0 and $\infty$. Thus, we will only discuss FPR convergence rates of relaxed PRS. The FPR identity

$$x^k_f - x^k_g = \frac{1}{2\lambda_k}(z^{k+1} - z^k),$$

shows that after $k$ iterations

$$\max\{d^2_{C_f}(x^k_f), d^2_{C_g}(x^k_g)\} \leq \|x^k_f - x^k_g\|^2 \overset{(3.1.13)}{=} o\left(\frac{1}{k+1}\right).$$

(3.7.3)

By the convexity of $C_f$ and $C_g$, the ergodic iterates of relaxed PRS satisfy $x^k_f = (1/\Lambda_k) \sum_{i=0}^k \lambda_i x^i_f \in C_f$ and $x^k_g = (1/\Lambda_k) \sum_{i=0}^k \lambda_i x^i_g \in C_g$. Thus, Theorem 2.5.1 of Chapter 2 implies the improved bound

$$\max\{d^2_{C_f}(\bar{x}_f^k), d^2_{C_g}(\bar{x}_g^k)\} \leq \|\bar{x}_f^k - \bar{x}_g^k\|^2 \leq O\left(\frac{1}{\Lambda_k^2}\right),$$

(3.7.4)

which is optimal by Theorem 2.7.1 of Chapter 2. Therefore, after $k$ iterations the relaxed PRS algorithm produces a point in each set with distance of order at most $O(1/\Lambda_k)$ from each other.
We now shift our focus to the MAP algorithm. First we replace both of the indicator functions with the squared distance functions:

\[ f(x) = \min_{y \in C_f} \|x - y\|^2 \quad \text{and} \quad g(x) = \min_{y \in C_g} \|x - y\|^2. \]

Now recall that \( f \) and \( g \) are differentiable, the gradient \( \nabla g \) is 2-Lipschitz continuous \cite[Corollary 12.30]{11}, and relaxed PRS takes the form in Equation (3.5.1). Specializing to \( \gamma = 1/2 \) and \( \lambda_k \equiv 1 \) yields the MAP algorithm (Corollary 3.5.1).

In this algorithm, the main MAP sequence satisfies \((z^j)_{j \geq 1} \subseteq C_f\), while the auxiliary sequences \((x^j_f)_{j \geq 0}\) and \((x^j_g)_{j \geq 0}\) are not necessarily elements \( C_f \) or \( C_g \). Therefore, the MAP FPR rate is less useful for estimating distances of the current iterates to \( C_f \) and \( C_g \) than it is in the relaxed PRS algorithm (See Equation (3.7.3)). Although \( \lambda_k \equiv 1 \), the map \( P_{C_f} P_{C_g} \) is \( \alpha \)-averaged for some \( \alpha < 1 \), and, hence, we can still estimate \( \|z^{k+1} - z^k\|^2 = o(1/(k + 1)) \) (Corollary 3.5.2).

The ergodic convergence rate in Theorem 2.5.1 of Chapter 2 (where we use the identity \( d_{C_f}(x^k_g) = (1/2)d_{C_g}(z^k) \) and Jensen’s inequality) shows that

\[
d_{C_g}^2 \left( \frac{1}{k + 1} \sum_{i=0}^{k} z^i \right) \leq \frac{2}{k + 1} \sum_{i=0}^{k} d_{C_g}^2 (x^i_f) = O \left( \frac{1}{k + 1} \right). \tag{3.7.5}
\]

Thus, if we choose \( z^0 \in C_f \), the ergodic iterate \((1/(k + 1)) \sum_{i=0}^{k} z^i \) is an element of \( C_f \) and we can bound its distance from \( C_g \). Note that this rate is strictly slower than the rate in Equation (3.7.4).

Although \( d_{C_f}^2 \) and \( d_{C_g}^2 \) are differentiable (Proposition 3.5.1), we cannot apply the results of Section 3.3 to MAP because they require that \((\lambda_j)_{j \geq 0} \subseteq (0, 1)\). Therefore, we cannot use the regularity of \( d_{C_f}^2 \) and \( d_{C_g}^2 \) to deduce faster convergence of the AP algorithm.

This discussion shows that the convergence rates predicted in Chapter 2 for relaxed PRS, which are known to be optimal, are faster than those predicted for MAP. When \( C_f \) and \( C_g \) intersect nicely (Section 3.5), the rate predicted for MAP is faster (See Corollary 3.5.1). In \cite[Section 8]{6} a similar phenomenon is observed for the case of intersecting
subspaces: DRS is faster than MAP for problems with nonregular intersection. It would be highly satisfying to characterize this phenomenon in general.

In the infeasible case, \( C_f \cap C_g = \emptyset \), the set of fixed points of \( T_{PRS} \) is empty and the iterates of the relaxed PRS algorithm are unbounded. On the other hand, under mild conditions (Remark 3.5.4), the iterates of MAP weakly converge to a point in \( C_f \), and \( P_{C_g} z^k - P_{C_f} P_{C_g} z^i \) strongly converges to the “gap vector” between the \( C_f \) and \( C_g \). This “best approximation” property is especially useful when we are unsure if \( C_f \cap C_g \neq \emptyset \).

3.7.2 Parallelized model fitting and classification

The following scenario appears in [26, Chapter 8]. Consider the model fitting problem:
Let \( M : \mathbb{R}^n \rightarrow \mathbb{R}^m \) be a feature matrix, let \( b \in \mathbb{R}^m \), be the output vector, let \( l \) be a loss function and let \( r \) be a regularization function. The goal of the model fitting problem is to

\[
\min_{x \in \mathbb{R}^n} l(Mx - b) + r(x). \tag{3.7.6}
\]

The function \( l \) is used to enforce the constraint \( Mx = b + \nu \) up to some noise \( \nu \) in the measurement, while \( r \) enforces the regularity of \( x \) by incorporating prior knowledge of the form of the solution.

In this section, we present several different ways to split Equation (3.7.6). Each splitting gives rise to a different algorithm and can be applied to general convex \( l \) and \( r \). Our discussion extends the one given in Section 2.9.2 of Chapter 2, where only convexity of \( l \) and \( r \) is assumed.

3.7.2.1 Auxiliary variable

We can split Equation (3.7.6) by defining an auxiliary variable for \( Mx - b \):

\[
\begin{align*}
\min_{x \in \mathbb{R}^n, y \in \mathbb{R}^m} & \quad l(y) + r(x) \\
\text{subject to} & \quad Mx - y = b. \tag{3.7.7}
\end{align*}
\]
We will now analyze the convergence rates predicted in Section 3.6.2 for ADMM applied to Problem (3.7.7). Our most general convergence result applies to the auxiliary terms:

\[
S_{dr}(w_k^d, w^*) = \max \left\{ \beta_r \frac{\alpha M}{2} \| w^*_k - w^* \|^2_{\mathbb{R}^m}, \frac{\mu_r}{2} \| Mx^k - Mx^* \|^2_{\mathbb{R}^m} \right\},
\]

\[
S_{dl}(w_k^d, w^*) = \max \left\{ \beta_l \frac{2}{\| w^*_k - w^* \|^2_{\mathbb{R}^m}}, \mu_l \frac{2}{\| y^*_k - y^* \|^2_{\mathbb{R}^n}} \right\}.
\]

Theorem 3.6.1 shows that the best auxiliary term converges with rate \( o(1/(k + 1)) \), the ergodic auxiliary term converges with rate \( O(1/\lambda_k) \), and the entire sequence of auxiliary terms converges with rate \( o(1/\sqrt{k+1}) \).

Now suppose that \( \mu_l > 0 \). Then we can bound the distance of \( y^k \) to the optimal point \( y^* := Mx^* - b \):

\[
\| y^k - y^* \|^2 = O \left( \frac{1}{\lambda_k^2} \right).
\]

Now let \( f = r \), let \( g = l \), let \( A = M \), and let \( B = -I_{\mathbb{R}^m} \). If \( \gamma < \kappa \mu_l \), then Theorem 3.6.2 bounds the primal objective error and the FPR:

\[
| l(y^k) + r(x^k) - l(Mx^* - b) - r(x^*) | = o \left( \frac{1}{k + 1} \right);
\]

\[
\| Mx^k - b - y^k \|^2 = o \left( \frac{1}{k + 1} \right).
\]

In particular, if \( l \) is Lipschitz, then \( | l(y^k) - l(Mx^k - b) | = o \left( 1/(k + 1) \right) \). Thus, we have

\[
0 \leq l(Mx^k - b) + r(x^k) - l(Mx^* - b) - r(x^*) = o \left( \frac{1}{k + 1} \right).
\]

A similar result holds if \( r \) is strongly convex and we assign \( g = r \) and \( f = l \), etc.

We can improve the above sublinear rate to a linear rate in any of the following cases (Theorem 3.6.3):

- \( r \) is differentiable and strongly convex and \( MM^* \) is strongly monotone;
- \( l \) is differentiable and strongly convex;
- \( r \) is differentiable, \( MM^* \) is strongly monotone, and \( l \) is strongly convex;
• $r$ is strongly convex and $l$ is differentiable.

In the following two splittings, we leave the derivation of convergence rates to the reader.

### 3.7.2.2 Splitting across samples

We assume that $l$ is block separable, i.e. $l(Mx - b) = \sum_{i=1}^{R} l_i(M_i x_i - b_i)$ where

$$M = \begin{bmatrix} M_1 \\ \vdots \\ M_R \end{bmatrix} \quad \text{ and } \quad b = \begin{bmatrix} b_1 \\ \vdots \\ b_R \end{bmatrix}.$$  

Each $M_i \in \mathbb{R}^{m_i \times n}$ is a submatrix of $M$, each $b_i \in \mathbb{R}^{m_i}$ is a subvector of $b$, and $\sum_{i=1}^{R} m_i = m$. Therefore, an equivalent form of Equation (3.7.6) is given by

$$\min_{x_1, \ldots, x_R, y \in \mathbb{R}^n} \sum_{i=1}^{R} l_i(M_i x_i - b_i) + r(y)$$

subject to $x_r - y = 0$, $r = 1, \ldots, R$. \hspace{1cm} (3.7.8)

We say that Equation (3.7.8) is **split across samples**. Thus, to apply ADMM to this problem, we simply stack the vectors $x_i$, $r = 1, \ldots, R$ into a vector $x = (x_1, \ldots, x_R)^T \in \mathbb{R}^{nR}$. Then the constraints in Equation (3.7.8) reduce to $Ax + By = 0$ where $A = I_{\mathbb{R}^{n \times R}}$ and $By = (-y, \ldots, -y)^T$.

### 3.7.2.3 Splitting across features

We can also split Equation (3.7.6) **across features** whenever $r$ is block separable in $x$, in the sense that there exists $C > 0$, such that for all $y \in \mathbb{R}^n$, $r(y) = \sum_{i=1}^{C} r_i(y_i)$ where $y_i \in \mathbb{R}^{n_i}$ and $\sum_{i=1}^{C} n_i = n$. This splitting corresponds to partitioning the columns of $M$, i.e.

$$M = \begin{bmatrix} M_1, \ldots, M_C \end{bmatrix}.$$
and $M_i \in \mathbb{R}^{m \times n_i}$ for all $i = 1, \cdots, C$. Note that for all $y \in \mathbb{R}^n$, the sum identity holds $My = \sum_{i=1}^CM_iy_i$. Now if we denote $x = (x_1, \cdots, x_C) \in \prod_{i=1}^C\mathbb{R}^m$, we can derive an equivalent form of Equation (3.7.8):

$$\minimize_{x \in \mathbb{R}^{mC}, y \in \mathbb{R}^n} \begin{cases} l \left( \sum_{i=1}^C x_i - b \right) + \sum_{i=1}^C r_i(y_i) \\
\text{subject to } x_i - M_i y_i = 0, \quad i = 1, \cdots, C. \end{cases} \quad \text{(3.7.9)}$$

The constraint in Equation (3.7.9) reduces to $Ax + By = 0$ where $A = I_{\mathbb{R}^{mc}}$ and $By = -(M_1 y_1, \cdots, M_C y_C)^T \in \mathbb{R}^{mc}$.

### 3.7.3 Linear and semidefinite programming

In this section we borrow the setting of [97]. The goal of linear (LP) and semidefinite (SDP) programming is to minimize a linear function subject to linear and matrix semidefinite constraints, respectively. Thus, in this section we study the following generic primal-dual pair problem

$$\begin{align*}
\minimize_{x \in \mathbb{R}^n} & \quad c^T x \\
\text{subject to } & \quad Ax + s = b \\
\maximize_{y \in \mathbb{R}^m} & \quad -b^T y \\
\text{subject to } & \quad -A^T y + r = c
\end{align*} \quad (3.7.10)$$

where $c \in \mathbb{R}^n$, $b, s \in \mathbb{R}^m$, $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear map, $K \subseteq \mathbb{R}^m$ is a closed convex cone, and $K^* \subseteq \mathbb{R}^m$ is the dual cone to $K$. In linear programming $K$, is the positive orthant $K = \mathbb{R}^n_+$, and for semidefinite programming, $K$ is the cone of symmetric, positive semidefinite matrices.

In [97], both optimization problems in Equation (1) are combined into a single feasibility problem. To this end we introduce slack variables $\tau, \kappa \in \mathbb{R}_+$, and the vectors and
matrix
\[
\begin{bmatrix}
x \\
y \\
\tau
\end{bmatrix} \in \mathbb{R}^{n+m+1}, \quad
\begin{bmatrix}
r \\
s \\
\kappa
\end{bmatrix} \in \mathbb{R}^{n+m+1}, \quad
\begin{bmatrix}
0 & A^T & c \\
-A & 0 & b \\
-c^T & -b^T & 0
\end{bmatrix} \in \mathbb{R}^{(n+m+1)\times(n+m+1)}.
\]

In addition, we let \(C = \mathbb{R}^n \times K^* \times \mathbb{R}_+\) and \(C^* = \{0\} \times K \times \mathbb{R}_+\). With this notation the goal of the \textit{homogeneous self dual embedding} problem is to find \((u, v) \in \mathbb{R}^{n+m+1}\) such that \(Qu = v\) and \((u, v) \in C \times C^*\).

Throughout this section we denote
\[
C_f = C \times C^* \quad \text{and} \quad C_g = \{(u, v) \in \mathbb{R}^{n+m+1} \times \mathbb{R}^{n+m+1} \mid Qu = v\}. \tag{3.7.11}
\]

Our goal is to find a point in the intersection \(C_f \cap C_g\). A remarkable trichotomy was derived in [115]: Suppose \((u, v) \in C_f \cap C_g\), then

1. If \(\tau > 0\) and \(\kappa = 0\), then \((x/\tau, y/\tau, s/\tau)\) is a primal dual solution of .

2. If \(\tau = 0\) and \(\kappa > 0\), then \(c^T x + b^T y < 0\). The case \(b^T y < 0\) is a certificate of primal infeasibility, and the case \(c^T x < 0\) is a certificate of dual infeasibility.

3. If \(\tau = \kappa = 0\), then nothing can be concluded about Equation (1). However, if there exists a point \((u', v') \in C_f \cap C_g\) for which \(\tau' + \kappa' \neq 0\), then we can choose an initial point \(z^0 \in \mathbb{R}^{n+m+1}\) such that DRS applied with \(f = \chi_{C_f}\) and \(g = \chi_{C_g}\) converges to a point \((u', v') \in C_f \cap C_g\) with \(\kappa' + \tau' \neq 0\).

3.7.3.1 Linear programming

Let us now examine the structure of the sets \(C_f\) and \(C_g\). For linear programming problems, \(C_f = \mathbb{R}^n \times \mathbb{R}_+^m \times \mathbb{R}_+ \times \{0\} \times \mathbb{R}_+^m \times \mathbb{R}_+\) is a polyhedron, i.e. the intersection of finitely many half planes, and \(C_g\) is a linear subspace. In finite dimensional spaces the pair \(\{C_f, C_g\}\) is linearly regular in the sense of Definition 3.5.1 [5, Remark 5.7.3].
We have four different algorithms that we can apply to find a point in $C_f \cap C_g$. The first two are the non parallelized versions of DRS which correspond to function pairs

\[(f = \chi_{C_f}, g = \chi_{C_g}) \quad \text{and} \quad (f = d_{C_f}^2, g = d_{C_g}^2). \quad (3.7.12)\]

Theorem 3.5.1 shows that relaxed PRS applied to the second pair (Equation (3.5.1)) linearly convergence to a point in the intersection $C_f \cap C_g$. Linear convergence of DRS applied to the first pair was shown in [13].

The projection onto $C_f$ is simple, and so the main computational bottleneck of the algorithm is to project onto $C_g$. There are various tricks that can be employed to speed this step up [97], but in some cases it is desirable to break up the linear equations into several sets $C_g = C_{g_1} \cap \cdots \cap C_{g_r}$, where $C_{g_i} \subseteq \mathbb{R}^{n+m+1}$ each encode a small number of linear constraints.

The collection $\{C_f, C_{g_1}, \cdots, C_{g_r}\}$ is linearly regular by [5, Remark 5.7.3], so we can apply Theorem 3.5.2 to show that

\[\{C_f \times C_{g_1} \times \cdots \times C_{g_r}, D\} \quad \text{is linearly regular} \]

where $D \subseteq \mathbb{R}^{(r+1)(n+m+1)}$ is the “diagonal set” of Section 3.5.1. Thus, we can apply DRS or relaxed PRS to either of the following pairs:

\[(f = \chi_{C_{f} \times C_{g_1} \times \cdots \times C_{g_r}}, g = \chi_D) \quad \text{and} \quad (f = d_{C_{f} \times C_{g_1} \times \cdots \times C_{g_r}}^2, g = d_D^2). \quad (3.7.13)\]

We can deduce linear convergence of the first pair using [13] and of the second by Theorem 3.5.3.

In general, the pairs in Equation (3.7.12) and (3.7.13) may not perform the same in practice. Thus, we cannot make any prediction about the practical performances of the methods. We can only point to our arguments in Section 3.7.1.2 that seem to indicate a better performance of the indicator function pair in problems that are badly conditioned.
3.7.3.2 Semidefinite programming

For semidefinite programming, \( \mathcal{K} \) is the cone of positive semidefinite matrices. Note that \( \mathcal{K}^* = \mathcal{K} \), i.e. \( \mathcal{K} \) is self dual [11, Example 6.25]. In general, the pair \( \{C_f, C_g\} \) is not necessarily (boundedly) linearly regular. The main condition to check is whether the relative interior of \( C_f \) intersects the subspace \( C_g \) [5, Theorem 5.6.2]. In fact, the relative interior of \( \mathcal{K} \) in \( \mathbb{R}^m \) is the set of all strictly positive definite matrices, i.e. the set of full rank positive definite matrices. Many problems of interest in semidefinite programming arise from the lifting of a non convex problem and desire low rank solutions of the associated SDP [66]. Thus, we do not expect the relative interior of \( C_f \) to intersect \( C_g \) for every SDP.

In terms of algorithm choice, we have at least four options to model the feasibility problem (See Equations (3.7.12) and (3.7.13)). In particular, when the linear constraints are difficult to solve in unison, we can break them into smaller pieces and solve them exactly. However, the main computational bottleneck of semidefinite programming is the projection onto the semidefinite cone. Unfortunately, there seems to be no way to lighten the cost of this projection.

The convergence rates for relaxed PRS applied to the feasibility problem are linear whenever the relative interior of \( C_f \) intersects \( C_g \). In terms of \( \mathcal{K} \) this condition requires that there is a full rank strictly positive definite primal dual pair \( (x, y) \in \mathcal{K} \times \mathcal{K} \). Finally, because we usually do not expect full rank solutions to SDPs, we just refer the reader to Section 3.7.1.2 and Equations (3.7.3) and (3.7.4) which show the worst case feasibility convergence rates.

3.8 Conclusion

In this chapter, we provided a comprehensive convergence rate analysis of relaxed PRS and ADMM under various regularity assumptions. By appealing to the examples developed in
Chapter 2, we showed that several of the convergence rates cannot be improved. All results follow from some combination of a lemma that deduces convergence rates of summable monotonic sequences (Lemma 3.1.1), a simple diagram (Figure 3.1), and fundamental inequalities (Propositions 3.1.3, 3.1.4, 3.3.1, and 3.3.2) that relate the FPR to the objective error of the relaxed PRS algorithm. Thus, together with Chapter 2, we have developed a comprehensive convergence rate of the relaxed PRS and ADMM algorithms under the standard regularity assumptions in convex optimization.
CHAPTER 4

Convergence Rate Analysis of Primal-Dual Splitting Schemes

4.1 Introduction

Primal-dual algorithms are abstract splitting schemes that solve monotone inclusion and convex optimization problems. These schemes fully decompose problems built from sums, linear compositions, parallel sums, and infimal convolutions of simple functions so that each simple term is processed individually. This decomposition is achieved by cleverly combining primal and dual pair problems into a single inclusion problem, to which standard operator splitting algorithms can be applied. This gives rise to algorithms that are inherently parallel or distributed and in which expensive matrix inversions can be avoided. The characteristics of primal-dual algorithms are especially desirable for large-scale applications in machine learning, image processing, distributed optimization, and control.

Primal-dual methods have a long history with many contributors, and an attempt to summarize and relate all of the contributions is beyond the scope of this chapter. In this chapter, we are mainly concerned with the line of work that began in [104, 36, 61] and the many generalizations and enhancements of the basic framework that followed [47, 52, 112, 29, 23, 24, 44, 78, 20, 45]. Thus, we consider the following prototypical convex optimization problem as our guiding example:

$$\min_{x \in \mathcal{H}_0} f(x) + g(x) + \sum_{i=1}^{n} (h_i \square l_i)(B_ix)$$  \hspace{1cm} (4.1.1)
where □ denotes the infimal convolution operation (see Section 4.1.2), \( n \in \mathbb{N} \), \( n \geq 1 \), \( \mathcal{H}_i \) are Hilbert spaces for \( i = 0, \cdots, n \), the functions \( f, g : \mathcal{H}_0 \rightarrow (-\infty, \infty] \) and \( h_i, l_i : \mathcal{H}_i \rightarrow (-\infty, \infty] \) are closed, proper, and convex for \( i = 1, \cdots, n \), and \( B_i : \mathcal{H}_0 \rightarrow \mathcal{H}_i \) is a bounded linear map for \( i = 1, \cdots n \).

All of the algorithms presented in this chapter completely disentangle the structure of Problem (4.1.1) so that each iteration only involves the individual proximal operators of each of the nondifferentiable terms, the gradient operators of the differentiable terms, and multiplication by the linear maps. Thus, the maps \( B_i \) are never inverted, and we never compute proximal operators or gradients of sums or infimal convolutions of functions. We note that this level of separability is not achieved by classical splitting methods such as forward-backward splitting, Douglas-Rachford splitting, or the alternating direction method of multipliers (ADMM) when they are applied directly to the primal optimization Problem (4.1.1) [32, 100, 65, 85].

In Problem (4.1.1), the maps \( B_i \) can be used as “data matrices,” in which case \( h_i \) and \( l_i \) are data fitting terms and \( f \) and \( g \) enforce prior knowledge on the structure of the solution, such as sparsity, low rank, or smoothness. In other cases, the maps \( h_i \) and \( l_i \) may be regularizers that emphasize many competing structures. We now present a particular example.

**Application: Constrained model fitting with group-structured regularizers.** Suppose we are given a measurement \( b \in \mathbb{R}^d \) and a dictionary \( A \in \mathbb{R}^{d \times m} \). Our goal is to recover a highly structured signal \( x = (x_1, \cdots, x_m) \in \mathbb{R}^m \) such that \( Ax \approx b \). For example, in the hierarchical sparse coding problem (HSCP) [74], we arrange the columns of \( A \) into a directed tree structure \( T \) and allow \( x_i = 0 \) only if \( x_j = 0 \) for all descendants \( j \) in \( T \) of node \( i \). Such a hierarchical representation is particularly useful for multi-scale data such as images and text documents. This type of regularization can be generalized to include arbitrary column groupings and complicated relationships between the elements of each group. Indeed, let \( G \) be a set of (possibly overlapping) subsets of \( \{1, \cdots, m\} \). For all \( S \in G \) and \( x \in \mathbb{R}^m \), let \( B_Sx = L_S(x_i)_{i \in S} \in \mathbb{R}^{m_S} \) where \( m_S > 0 \) and \( L_S : \mathbb{R}^{|S|} \rightarrow \mathbb{R}^{m_S} \) is a
linear map. Let $C \subseteq \mathbb{R}^m$ be a closed convex set, and let $\iota_C$ be the convex indicator function of $C$. For all $S \in G$, let $h_S : \mathbb{R}^{m_S} \to (-\infty, \infty]$ be a regularizer, and let $l_S = \iota_{\{0\}}$, which implies $h_S \sqcap l_S = h_S$. Then one special case of Problem (4.1.1) is the group-structured regularized model fitting problem:

$$\minimize_{x \in \mathbb{R}^m} \iota_C(x) + (1/2)\|Ax - b\|^2 + \sum_{S \in G} h_S(B_S x).$$

In [74], the authors consider the nonegativity constraint $C = \mathbb{R}^m_{\geq 0}$ and a grouping $G$ which consists of overlapping sets $S_i$ for $i \in \{1, \ldots, m\}$ such that $S_i$ contains $i$ and all of the descendants of $i$ in $\mathcal{T}$. Furthermore, for each $S \in G$, they consider the map $L_S = I_{\mathbb{R}^{\mid S\mid}}$ and the function $h_S = w_S \| (x_i)_{i \in S} \|_p$ where $p \in [1, \infty]$ and $w_S > 0$. This setup induces a mixed $\ell_1/\ell_p$ norm on $\mathbb{R}^m$ of the form $\sum_{S \in G} w_S \| (x_i)_{i \in S} \|_p$, which tends to “zero out” entire groups of components. Note that the sum is also highly nonseparable in the components of $x$, which can make the proximal operator of the regularization term difficult to evaluate.

If we denote $f(x) = \iota_C(x)$ and $g(x) = (1/2)\|Ax - b\|^2$, then the algorithms in this chapter only utilize the projection $P_C = \text{prox}_f$ onto $C$, the gradient $\nabla g(x) = A^*(Ax - b)$, and for all $S \in G$ in parallel, multiplications by the maps $B_S$ and $B_S^*$, and evaluations of the proximal operator of the function $h_S$. Not only does this make each iteration of the algorithm simple to implement and computationally inexpensive, it also provides a unified algorithmic framework for higher order regularizations of the components in each group, a task which might otherwise be intractable in large-scale applications.

Finally, we note that the use of infimal convolutions in applications is not wide-spread, so we list a few instances where they may be useful: Infimal convolutions are used in image recovery [35, Section 5] to remove staircasing effects in the total variation model. The infimal convolution of the indicator functions of two closed convex sets is the indicator function of their Minkowski sum, which has applications in motion planning for robotics [80, Section 4.3.2]. In convex analysis, the Moreau envelope of a function arises as an infimal convolution with a multiple of the squared norm [11, Section 12.4]. More generally, the infimal convolution of $h_i$ and $l_i$ can be interpreted as a regularization or smoothing of $h_i$ by $l_i$ and vice versa [11, Section 18.3].
4.1.1 Goals, challenges, and approaches

This work seeks to improve the theoretical understanding of the convergence rates of primal-dual splitting schemes. In this chapter, we study primal-dual algorithms that are applications of standard operator splitting algorithms in product spaces consisting of primal and dual variables. Consequently, the convergence theory for these algorithms is well-developed, and they are known to converge (weakly) under mild conditions.

Although we understand when these algorithms converge, relatively little is known about their rate of convergence. For convex optimization algorithms, the ergodic convergence rate of the primal-dual gap has been analyzed in a few cases [36, 21, 20, 109]. However, even in cases where convergence rates are known, variable metrics and stepsizes, which can significantly improve practical performance of the algorithms [103, 67], are not analyzed. In addition, we are not aware of any convergence rate analysis of the primal-dual gap for the nonergodic (or last) iterate generated by these algorithms. It is important to understand nonergodic convergence rates because the ergodic (or time-averaged) iterates can “average out” structural properties, such as sparsity and low rank, that are shared by the solution and the nonergodic iterate.

The convergence rate analysis of the ergodic primal-dual gap largely follows from subgradient inequalities and an application of Jensen’s inequality. In contrast, the techniques developed in this chapter exploit the properties of the nonexpansive operators driving the algorithms to deduce the nonergodic convergence rate of the primal-dual gap. Thus, our techniques are quite different from those used in classical convergence rate analysis and parallel the analysis developed in Chapter 2.

We summarize our contributions and techniques as follows:

i We describe a model monotone inclusion problem that generalizes many primal-dual formulations that appear in the literature. We provide a simple prototype algorithm to solve the model problem, and we deduce a fundamental inequality that bounds the primal-dual gap at each iteration of the algorithm. We then simplify the inequality in
the special case of four splitting algorithms (Section 4.2).

ii We derive ergodic convergence rates of the variable metric forms of the relaxed proximal point algorithm (PPA), relaxed forward-backward splitting (FBS), and forward-backward-forward splitting as well as the fixed metric relaxed Peaceman-Rachford splitting (PRS) algorithm (Section 4.3). After some algebraic simplifications, our analysis essentially follows from an application of Jensen’s inequality.

iii We derive nonergodic convergence rates of relaxed PPA, relaxed FBS, and relaxed PRS (Section 4.4). All of our analysis follows by bounding the primal-dual gap function by a multiple of the fixed-point residual (FPR) of the nonexpansive mapping that drives the algorithm. Thus, we show that the size of the FPR can be used as a valid stopping criteria for these three algorithms.

iv We apply our results to deduce ergodic and nonergodic convergence rates for a large class of primal-dual algorithms that have appeared in the literature (Section 4.5).

Our analysis not only deduces the convergence rates of a large class of primal-dual algorithms found in the literature. It also serves as a resource for the analysis of future primal-dual algorithms that solve generalizations of Problem 4.1.1, e.g. [15, 24].

4.1.2 Definitions, notation and some facts

In what follows, $\mathcal{H}, \mathcal{G}$, and $\mathbf{H}$ denote (possibly infinite dimensional) Hilbert spaces. We always use the notation $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ to denote the inner product and norm associated to a Hilbert space, respectively. Note that there is some ambiguity in this convention, but it simplifies the notation and no confusion should arise. The space $\mathbf{H}$ will usually denote a product Hilbert space consisting of primal variables in $\mathcal{H}$ and dual variables in $\mathcal{G}$. Let $\mathbb{R}_{++} = \{ x \in \mathbb{R} \mid x > 0 \}$ denote the set of strictly positive real numbers. In all of the algorithms we consider, we utilize two stepsize sequences: the implicit sequence $(\gamma_j)_{j \geq 0} \subseteq \mathbb{R}_{++}$ and the explicit sequence $(\lambda_j)_{j \geq 0} \subseteq \mathbb{R}_{++}$. We define the $k$-th partial sum
of the sequence $(\gamma_j \lambda_j)_{j \geq 0}$ by the formula:

$$\Sigma_k := \sum_{i=0}^{k} \gamma_i \lambda_i.$$  \hfill (4.1.2)

Given a sequence $(x^j)_{j \geq 0} \subset \mathcal{H}$, we let $\overline{x}^k = (1/\Sigma_k) \sum_{i=0}^{k} \gamma_i \lambda_i x^i$ denote its $k$th average with respect to the sequence $(\gamma_j \lambda_j)_{j \geq 0}$. We call a convergence result \textit{ergodic} if it is in terms of the sequence $(\overline{x}^j)_{j \geq 0}$, and \textit{nonergodic} if it is in terms of $(x^j)_{j \geq 0}$.

We denote the set of summable nonnegative sequences by $\ell_1^+(\mathbb{N}) := \{(\eta_j)_{j \geq 0} \subseteq [0, \infty) | \sum_{j=0}^{\infty} \eta_j < \infty\}$.

The following definitions and facts are mostly standard and can be found in [11, 48].

We let $\mathcal{B}(\mathcal{H}, \mathcal{G})$ denote the set of bounded linear maps from $\mathcal{H}$ to $\mathcal{G}$, and set $\mathcal{B}(\mathcal{H}) := \mathcal{B}(\mathcal{H}, \mathcal{H})$. We will use the notation $I_{\mathcal{H}} \in \mathcal{B}(\mathcal{H})$ to denote the identity map. Given a map $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$, we denote its adjoint by $L^* \in \mathcal{B}(\mathcal{G}, \mathcal{H})$. The operator norm on $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$ is defined by the following supremum: $\|L\| = \sup_{x \in \mathcal{H}, \|x\| \leq 1} \|Lx\|$. Let $\rho \in \mathbb{R}$ be a nonnegative real number. We let $\mathcal{S}_\rho(\mathcal{H}) \subseteq \mathcal{B}(\mathcal{H})$ denote the set of linear $\rho$-strongly monotone self-adjoint maps:

$$\mathcal{S}_\rho(\mathcal{H}) := \{U \in \mathcal{B}(\mathcal{H}) | U = U^*, (\forall x \in \mathcal{H}) \langle Ux, x \rangle \geq \rho \|x\|^2\}.$$

We define the norm and inner product induced by $U \in \mathcal{S}_\rho(\mathcal{H})$ on $\mathcal{H}$ by the formulae: for all $x, y \in \mathcal{H}$, $\|x\|_U^2 := \langle Ux, x \rangle$, and $\langle x, y \rangle_U = \langle Ux, y \rangle$. The Loewner partial ordering on $\mathcal{S}_\rho(\mathcal{H})$ is as follows:

$$U_1 \succeq U_2 \iff (\forall x \in \mathcal{H}) \|x\|_{U_1}^2 \geq \|x\|_{U_2}^2.$$

Let $L \geq 0$, and let $D$ be a nonempty subset of $\mathcal{H}$. A map $T : D \to \mathcal{H}$ is called $L$-Lipschitz if for all $x, y \in \mathcal{H}$, we have $\|Tx - Ty\| \leq L \|x - y\|$. In particular, $T$ is called \textit{nonexpansive} if it is 1-Lipschitz. A map $N : D \to \mathcal{H}$ is called $\lambda$-averaged [11, Section 4.4] if

$$N = T_{\lambda} := (1 - \lambda)I_{\mathcal{H}} + \lambda T$$ \hfill (4.1.3)
for a nonexpansive map $T : D \to \mathcal{H}$ and a real number $\lambda \in (0, 1)$. A $(1/2)$-averaged map is called \textit{firmly nonexpansive}. We will always use a $^*$ superscript to denote a fixed point of a nonexpansive map, e.g., $z^*$.

Let $2^\mathcal{H}$ denote the power set of $\mathcal{H}$. A set-valued operator $A : \mathcal{H} \to 2^\mathcal{H}$ is called \textit{monotone} if for all $x, y \in \mathcal{H}$, $u \in Ax$, and $v \in Ay$, we have $\langle x - y, u - v \rangle \geq 0$. We denote the set of zeros of a monotone operator by $\text{zer}(A) := \{ x \in \mathcal{H} \mid 0 \in Ax \}$. The \textit{graph} of $A$ is denoted by $\text{gra}(A) := \{ (x, y) \mid x \in \mathcal{H}, y \in Ax \}$. Evidently, $A$ is uniquely determined by its graph. A monotone operator $A$ is called \textit{maximal monotone} provided that $\text{gra}(A)$ is not properly contained in the graph of any other monotone set-valued operator. The \textit{inverse} of $A$, denoted by $A^{-1}$, is defined uniquely by its graph: $\text{gra}(A^{-1}) := \{ (y, x) \mid x \in \mathcal{H}, y \in Ax \}$. Let $\beta \in \mathbb{R}^{++}$ be a positive real number. The operator $A$ is called $\beta$-\textit{strongly monotone} provided that for all $x, y \in \mathcal{H}$, $u \in Ax$, and $v \in Ay$, we have $\langle x - y, u - v \rangle \geq \beta \| x - y \|^2$. A single-valued operator $B : \mathcal{H} \to 2^\mathcal{H}$ maps each point in $\mathcal{H}$ to a singleton and will be identified with the natural $\mathcal{H}$-valued map it defines. A single-valued operator $B$ is called $\beta$-\textit{cocoercive} provided that for all $x, y \in \mathcal{H}$, we have $\langle x - y, Bx - By \rangle \geq \beta \| Bx - By \|^2$. Evidently, $B$ is $\beta$-cocoercive whenever $B^{-1}$ is $\beta$-strongly monotone. The parallel sum of (not necessarily single-valued) monotone operators $A$ and $B$ is given by $A \square B := (A^{-1} + B^{-1})^{-1}$. The \textit{resolvent} of a monotone operator $A$ is defined by the inversion $J_A := (I + A)^{-1}$. Minty’s theorem shows that $J_A$ is single-valued and has full domain $\mathcal{H}$ if, and only if, $A$ is maximally monotone. Note that $A$ is monotone if, and only if, $J_A$ is firmly nonexpansive. Thus, the \textit{reflection operator}

$$\text{refl}_A := 2J_A - I_{\mathcal{H}}$$

is nonexpansive on $\mathcal{H}$ whenever $A$ is maximally monotone. If $\rho > 0$ and $U \in \mathcal{S}_\rho(\mathcal{H})$, the operator $U^{-1}A$ is maximal monotone in $\langle \rangle_U$, if, and only if, $A$ is maximally monotone in $\langle \rangle$. Let $\gamma \in (0, \infty)$. The resolvent of the map $\gamma U^{-1}A$ has the special identity: $J_{\gamma U^{-1}A} = U^{-1/2}J_{\gamma U^{-1/2}AU^{-1/2}U^{1/2}}$ [49, Example 3.9].

Let $\Gamma_0(\mathcal{H})$ denote the set of closed, proper, and convex functions $f : \mathcal{H} \to (-\infty, \infty]$. 139
Let \( \text{dom}(f) := \{ x \in \mathcal{H} \mid f(x) < \infty \} \). We will let \( \partial f(x) : \mathcal{H} \to 2^\mathcal{H} \) denote the subdifferential of \( f \): \( \partial f(x) := \{ u \in \mathcal{H} \mid \forall y \in \mathcal{H}, f(y) \geq f(x) + \langle y - x, u \rangle \} \). We will always let
\[
\tilde{\nabla} f(x) \in \partial f(x)
\] (4.1.5)
denote a subgradient of \( f \) drawn at the point \( x \), and the actual choice of the subgradient \( \tilde{\nabla} f(x) \) will always be clear from the context; note that this notation was also used in [17].

The subdifferential operator of \( f \) is maximally monotone. The inverse of \( \partial f \) is given by \( \partial f^* \) where \( f^*(y) := \sup_{x \in \mathcal{H}} \langle y, x \rangle - f(x) \) is the Fenchel conjugate of \( f \). If the function \( f \) is \( \beta \)-strongly convex, then \( \partial f \) is \( \beta \)-strongly monotone.

If a convex function \( f : \mathcal{H} \to (-\infty, \infty] \) is Fréchet differentiable at \( x \in \mathcal{H} \), then \( \partial f(x) = \{ \nabla f(x) \} \). Suppose \( f \) is convex and Fréchet differentiable on \( \mathcal{H} \), and let \( \beta \in \mathbb{R}_{++} \) be a positive real number. Then the Baillon-Haddad theorem states that \( \nabla f \) is \( (1/\beta) \)-Lipschitz, if, and only if, \( \nabla f \) is \( \beta \)-cocoercive.

The resolvent operator associated to \( \partial f \) is called the \emph{proximal operator} and is uniquely defined by the following (strongly convex) minimization problem:
\[
\text{prox}_f(x) := J_{\partial f}(x) = \arg\min_{y \in \mathcal{H}} f(y) + (1/2)\|y - x\|^2.
\] If \( \rho > 0 \), \( U \in \mathcal{S}_{\rho}(\mathcal{H}) \), and \( \gamma \in (0, \infty) \), the proximal operator of \( f \) in the metric induced by \( U \) is given by the following formula: for all \( x \in \mathcal{H} \),
\[
\text{prox}_U^{\gamma f}(x) := J_{\gamma U^{-1}\partial f}(x) = \arg\min_{y \in \mathcal{H}} f(y) + \frac{1}{2\gamma} \|y - x\|_U^2.
\] (4.1.6)

The \emph{infimal convolution} of two functions \( f, g : \mathcal{H} \to (-\infty, \infty] \) is denoted by \( f \Box g : \mathcal{H} \to [-\infty, \infty] : x \mapsto \inf_{y \in \mathcal{H}} f(y) + g(x - y) \). The indicator function of a closed, convex set \( C \subseteq \mathcal{H} \) is denoted by \( \iota_C : \mathcal{H} \to \{0, \infty\} \); the indicator function is 0 on \( C \) and is \( \infty \) on \( \mathcal{H} \setminus C \).

Finally, we call the following identity the \emph{cosine rule}:
\[
\|y - z\|^2 + 2\langle y - x, z - x \rangle = \|y - x\|^2 + \|z - x\|^2, \quad \forall x, y, z \in \mathcal{H}.
\] (4.1.7)
4.1.3 Assumptions

Assumption 4.1.1 (Convexity). *Every function we consider is closed, proper, and convex.*

Unless otherwise stated, a function is not necessarily differentiable.

Assumption 4.1.2 (Differentiability). *Every differentiable function we consider is Fréchet differentiable* [11, Definition 2.45].

We employ other assumptions throughout the chapter, but we list them closer to where they are invoked.

4.1.4 Basic properties of metrics

A simple proof of the following Lemma recently appeared in [48, Lemma 2.1]. It previously appeared in [75, Section VI.2.6].

Lemma 4.1.1 (Metric properties). *Whenever* $U, V \in S_0(\mathcal{H})$ *satisfy the inequality* $\alpha I_{\mathcal{H}} \succeq U \succeq V \succeq \beta I_{\mathcal{H}}$ *for* $\alpha, \beta > 0$, *we have the ordering* $(1/\beta)I_{\mathcal{H}} \succeq V^{-1} \succeq U^{-1} \succeq (1/\alpha)I_{\mathcal{H}}$, *the inclusion* $U^{-1} \in S_{\|U\|-1}(\mathcal{H})$, *and the inequality* $\|U^{-1}\| \leq (1/\beta)$.

4.1.5 Basic properties of resolvents and averaged operators

The following are simple modifications of standard facts found in [11].

Proposition 4.1.1. *Let* $\rho > 0$, *let* $\lambda > 0$, *let* $\alpha \in (0, 1)$, *let* $U \in S_\rho(\mathcal{H})$, *let* $A : \mathcal{H} \to \mathcal{H}$ *be a single-valued maximal monotone operator*, and let $f \in \Gamma_0(\mathcal{H})$

1. **Optimality conditions of J**: *We have* $x^+ := J_{\gamma U^{-1}(\partial f + A)}(x)$ *if, and only if, there exists a unique subgradient* $\nabla f(x^+) := (1/\gamma)U(x - x^+) - Ax^+ \in \partial f(x^+)$, *such that* 

$$
\nabla f(x^+) + Ax^+ = \frac{1}{\gamma}U(x - x^+) \in \partial f(x^+) + Ax^+.
$$
2. **Averaged operator contraction property:** Let $\lambda \in (0, 1)$. A map $T : \mathcal{H} \to \mathcal{H}$ is $\lambda$-averaged in the metric induced by $U$ if, and only if, for all $x, y \in \mathcal{H}$

$$\|Tx - Ty\|_U^2 \leq \|x - y\|_U^2 - \frac{1 - \lambda}{\lambda} \|(I_H - T)x - (I_H - T)y\|_U^2. \quad (4.1.8)$$

3. **Wider relaxations:** A map $T : \mathcal{H} \to \mathcal{H}$ is $\alpha$-averaged in $\|\cdot\|_U$, if, and only if, $T_{\lambda}$ (Equation (4.1.3)) is $\lambda\alpha$-averaged in $\|\cdot\|_U$ for all $\lambda \in (0, 1/\alpha)$. In addition, $T_{1/\alpha}$ is nonexpansive with respect to $\|\cdot\|_U$.

### 4.1.6 Variable metrics

Throughout this chapter we will consider sequences of mappings $(U_j)_{j \geq 0} \in S_\rho(\mathcal{H})$ for some $\rho > 0$. In order to apply the standard convergence theory for variable metrics, we will make the following assumption:

**Assumption 4.1.3.** There exists a summable sequence $(\eta_j)_{j \geq 0} \subseteq \ell_1^+(\mathbb{N})$ such that for all $k \geq 0$, $1 + \eta_k U_k \succeq U_{k+1}$. In addition $\mu := \sup_{j \geq 0} \|U_j\| < \infty$.

Assumption 4.1.3 is standard in variable metric algorithms [48, 113, 49, 99].

**Remark 4.1.1.** There is an asymmetry in our notation and the notation of [48, 113, 49, 99]. In our analysis, the map $U \in S_\rho(\mathcal{H})$ induces a metric on $\mathcal{H}$. In other papers, the maps $U^{-1}$ induce a metric on $\mathcal{H}$.

The following notation will be used throughout the rest of the chapter. The proof is elementary.

**Proposition 4.1.2** (Metric parameters). Suppose that $(\eta_j)_{j \geq 0} \subseteq \ell_1^+(\mathbb{N})$. Define

$$\eta_p := \prod_{i=0}^{\infty} (1 + \eta_i) \quad \text{and} \quad \eta_s := \sum_{i=0}^{\infty} \eta_i.$$ 

Then $\eta_p$ and $\eta_s$ are finite.

The following Proposition is a consequence of the proof of [48, Theorem 5.1]. The proof is simple, so we omit it.
Proposition 4.1.3. Let $\mathcal{H}$ be a Hilbert space. Let $\rho \in (0, \infty)$, let $(\eta_j)_{j \geq 0} \subseteq \ell_1^1(\mathbb{N})$, and let $(U_j)_{j \geq 0} \in \mathcal{S}_\rho(\mathcal{H})$ satisfy Assumption 4.1.3. For all $k \geq 0$, let $\alpha_k \in (0,1)$, let $\lambda_k \in (0,1/\alpha_k]$ be a relaxation parameter, and let $T_k : \mathcal{H} \to \mathcal{H}$ be $\alpha_k$-averaged in the metric $\| \cdot \|_{U_k}$. Furthermore, assume that there is a point $z^* \in \mathcal{H}$ such that $T_k z^* = z^*$ for all $k \geq 0$. Let the $(z^j)_{j \geq 0}$ be generated by the following Krasnosel’ski˘ı-Mann (KM)-type iteration (Equation (4.1.3)):

$$z^{k+1} = (T_k)_{\lambda_k} z^k.$$ 

Then the following are true:

1. For all $k \geq 0$, $\|z^{k+1} - z^*\|_{U_k}^2 \leq (1 + \eta_k) \|z^k - z^*\|_{U_k}^2$ and, hence,

$$\|z^k - z^*\|_{U_k}^2 \leq \eta_p \|z^0 - z^*\|_{U_0}^2.$$

2. The following sum is finite:

$$\sum_{i=0}^{\infty} \frac{1 - \alpha_i \lambda_i}{\alpha_i \lambda_i} \|z^{i+1} - z^i\|^2 \leq \frac{1}{\rho} (1 + \eta_p \eta_k) \|z^0 - z^*\|_{U_0}^2.$$

We will use the following proposition to select parameters in the FBS algorithm. The proof of the following fact follows from [112, Equation (3.35)]

Proposition 4.1.4. Let $\rho > 0$, let $B : \mathcal{H} \to \mathcal{H}$ be $\beta$-cocoercive in the norm $\| \cdot \|$, and let $U \in \mathcal{S}_\rho(\mathcal{H})$. Then $U^{-1} B$ is $\beta \rho$-cocoercive in the norm $\| \cdot \|_U$.

The following proposition essentially follows from the proof of [113, Theorem 3.1].

Proposition 4.1.5. Let $A : \mathcal{H} \to 2^\mathcal{H}$ be maximal monotone, let $B : \mathcal{H} \to \mathcal{H}$ be monotone and $(1/\beta)$-Lipschitz, let $\rho > 0$, let $(U_j)_{j \geq 0} \subseteq \mathcal{S}_\rho(\mathcal{H})$ satisfy Assumption 4.1.3, and let $(\gamma_j)_{j \geq 0} \subseteq (0, \rho \beta]$. Let $(z^j)_{j \geq 0}$ be a sequence of points defined by the iteration: for all $k \geq 0$,

$$y^k = z^k - \gamma_k U_k^{-1} B z^k;$$

$$x^k = J_{\gamma_k U_k^{-1} A}(y^k);$$

$$w^k = x^k - \gamma_k U_k^{-1} B x^k;$$

$$z^{k+1} = z^k - y^k + w^k.$$
Suppose that \( \text{zer}(A + B) \neq \emptyset \). Then for all \( z^* \in \text{zer}(A + B) \) and for all \( k \geq 0 \), we have:
\[
\| z^{k+1} - z^* \|^2_{U_{k+1}} \leq (1 + \eta_k) \| z^k - z^* \|^2_{U_k}.
\]

4.2 The unifying scheme

In this section, we introduce a prototype monotone inclusion problem that generalizes and summarizes many primal-dual problem formulations found in the literature. After we describe the problem, we will introduce an abstract unifying scheme that generalizes many existing primal-dual algorithms. We will describe how to measure convergence of the unifying scheme, and introduce a fundamental inequality that bounds our measure of convergence. Finally, we will identify the key terms in the fundamental inequality and simplify them in the case of several abstract splitting algorithms.

In Section 4.5, we will show that this unifying scheme relates to many existing algorithms, and extend the convergence rate results of those methods.

4.2.1 Problem and algorithm

We focus on the following problem:

**Problem 4.2.1** (Prototype primal-dual problem). Let \((H, \langle \cdot, \cdot \rangle)\) be a Hilbert space, let \( f, g \in \Gamma_0(H) \), and let \( S : H \to H \) be a skew symmetric map: \( S^* = -S \). Then the prototype primal-dual problem is to find \( x^* \in H \) such that

\[
0 \in \partial f(x^*) + \partial g(x^*) + Sx^*.
\]

Evidently, Problem 4.2.1 is a monotone inclusion problem because \( \partial f, \partial g, \) and \( S \) are maximally monotone operators on \( H \) [11, Example 20.30].

144
We are now ready to define our unifying scheme.

<table>
<thead>
<tr>
<th>Algorithm 6: Unifying scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>input</strong>: $z^0 \in H, (\lambda_j)<em>{j \geq 0} \subseteq \mathbb{R}^{++}, (\gamma_j)</em>{j \geq 0} \subseteq \mathbb{R}^{++}, \rho &gt; 0, (U_j)<em>{j \geq 0} \subseteq S</em>\rho(H)$</td>
</tr>
<tr>
<td><strong>for</strong> $k = 0, 1, \ldots$ <strong>do</strong></td>
</tr>
<tr>
<td>$z^{k+1} = z^k - \gamma_k \lambda_k U^{-1}_k \left( \tilde{\nabla} f(x^k_f) + \tilde{\nabla} g(x^k_g) + S x^k_S \right)$;</td>
</tr>
</tbody>
</table>

Note that the points $x^k_f, x^k_g$, and $x^k_S$ as well as the subgradients $\tilde{\nabla} f(x^k_f) \in \partial f(x^k_f)$ and $\tilde{\nabla} g(x^k_g) \in \partial g(x^k_g)$ are unspecified in the description of Algorithm 6. In the algorithms we study, these points and subgradients will be generated by proximal and forward gradient operators and, thus, can be determined given $z^k$; see Section 4.2.2 for examples. However, Algorithm 6 is only meant to illustrate the algebraic form that our analysis addresses, and it is not meant to be an actual algorithm that solves Problem 4.2.1. The positive scalar sequence $(\lambda_j)_{j \geq 0}$ consists of relaxation parameters, or explicit steplike parameters, whereas the sequence $(\gamma_j)_{j \geq 0}$ consists of proximal parameters, or implicit steepest parameters. The strongly monotone maps $(U_j)_{j \geq 0}$ induce the metrics used in each iteration of the algorithm.

In all of our applications, $H$ will be a product space of primal and dual variables. In this setting, $f$ and $g$ will be block-separable maps, and $g$ will sometimes be differentiable. The map $S$ “mixes” the primal and dual variable sequences in the product space. Mixing is necessary, because the sequences are otherwise uncoupled.

The sequence of maps $(U_j)_{j \geq 0}$ is employed for two purposes. First, the maps are used because the evaluation of the resolvent $J_{\partial f^* + S}$, which is a basic building block of most of the algorithms we study, may not be simple. Thus, the primal-dual algorithms that we study formulate special metrics induced by $U \in S_\rho(H)$ such that $J_{U^{-1}(\partial f^* + S)}$ is as easy to evaluate as $\text{prox}_f$ (See Section 4.5). Hence, in our analysis we must at least consider fixed metrics that are different from the standard product metric on $H$. Second, we allow the metrics to vary at each iteration because it can significantly improve the practical performance of the algorithm, e.g. by employing second order information, or even simple time-varying diagonal metrics [103, 67].
4.2.2 Examples of the unifying scheme

In this section we introduce four algorithms and show that they are a special case of Algorithm 6. We will also introduce several assumptions on the algorithm parameters that ensure convergence. These assumptions will remain in effect throughout the rest of the chapter. Note that the convergence theory of the methods in this section is well-studied. See [11, 48, 112, 49, 106, 111, 85] for background. Finally, we will say that several algorithms in this section are relaxed. For brevity, we will drop this adjective throughout whenever convenient.

The relaxed variable metric PPA applies to problems in which $g \equiv 0$.

**Algorithm 7:** Relaxed variable metric proximal point algorithm (PPA)

**input:** $z^0 \in H; (\lambda_j)_{j \geq 0} \subseteq (0, 2]; (\gamma_j)_{j \geq 0} \subseteq \mathbb{R}^{++}; \rho > 0; (U_j)_{j \geq 0} \subseteq \mathcal{S}_\rho(H);

**for** $k = 0, 1, \ldots$ **do**

\[
\begin{aligned}
x^k_g &= z^k; \\
x^k_f &= J_{\gamma_k U_k^{-1}(\partial I + S)}(z^k); \\
z^{k+1} &= (1 - \lambda_k)z^k + \lambda_k x^k_f;
\end{aligned}
\]

The relaxed variable metric FBS algorithm can be applied whenever $g$ is differentiable
and $\nabla g$ is $(1/\beta)$-Lipschitz for some $\beta > 0$.

**Algorithm 8:** Relaxed variable metric forward-backward algorithm (FBS)

**input:** $z^0 \in H; \rho > 0; \varepsilon \in (0, 2\beta \rho);$ 
$(\gamma_j)_{j \geq 0} \subseteq (0, 2\beta \rho - \varepsilon);$ 
$\alpha_k := (2\beta \rho)/(4\beta \rho - \gamma_k)$ for $k \geq 0;$ 
$\delta \in (0, \min\{1/\alpha_j \mid j \geq 0\});$ 
$\lambda_k \in (0, 1/\alpha_k - \delta]$ for $k \geq 0;$ 
$(U_j)_{j \geq 0} \subseteq S_\rho(H);$ 

for $k = 0, 1, \ldots$ do 

$\begin{align*}
  x^k_g &= z^k; \\
  x^k_f &= J_{\gamma_k U_k}^{-1}(\partial f + S)(z^k - \gamma_k U_k^{-1} \nabla g(z^k)); \\
  z^{k+1} &= (1 - \lambda_k)z^k + \lambda_k x^k_f;
\end{align*}$

In the relaxed PRS algorithm, we fix the metric and the implicit stepsize parameters throughout the course of the algorithm. We do this because the fixed-points of the PRS operator can vary with $\gamma$ and $U$. Thus, changing these parameters will lead to an algorithm that “chases” a new fixed-point at each iteration.

**Algorithm 9:** Relaxed Peaceman-Rachford splitting (PRS)

**input:** $z^0 \in H; (\lambda_j)_{j \geq 0} \subseteq (0, 2]; \gamma > 0; \rho > 0; U \in S_\rho(H); w \in R;$ 

for $k = 0, 1, \ldots$ do 

$\begin{align*}
  z^{k+1} &= (1 - \frac{\lambda_k}{2})z^k + \frac{\lambda_k}{2} \text{refl}_{\gamma U^{-1}(\partial f + w S)} \circ \text{refl}_{\gamma U^{-1}(\partial g + (1-w)S)}(z^k);
\end{align*}$

The variable metric FBF algorithm can be applied whenever $g$ is differentiable and
∇g is \((1/\beta)\)-Lipschitz for some \(\beta > 0\).

**Algorithm 10:** Variable metric forward-backward-forward algorithm (FBF)

**input:** \(z^0 \in H; \rho > 0; (U_j)_{j \geq 0} \subseteq S_\rho(H); (\gamma_j)_{j \geq 0} \subseteq (0, \rho/(\beta^{-1} + \|S\|))\);

for \(k = 0, 1, \ldots\) do

\[
y^k = z^k - \gamma_k U_k^{-1}(\nabla g(z^k) + S z^k);
\]

\[
x_f^k = J_{\gamma_k U_k^{-1}} \partial f(y^k);
\]

\[
w^k = x_f^k - \gamma_k U_k^{-1}(\nabla g(x_f^k) + S x_f^k);
\]

\[
z^{k+1} = z^k - y^k + w^k;
\]

The following lemma relates the above algorithms to the unifying scheme.

**Lemma 4.2.1.** Algorithms 7, 8, 9, and 10 are special cases of the unifying scheme. In particular, the following hold:

1. In Algorithm 7, we have \(x_g^k := z^k, x_S^k := x_f^k\), and

\[
\nabla f(x_f^k) := (1/\gamma_k)U_k(z^k - x_f^k) - S x_f^k \in \partial f(x_f^k)
\]

for all \(k \geq 0\).

2. In Algorithm 8, we have \(x_g^k := z^k, x_S^k := x_f^k\), and

\[
\nabla f(x_f^k) := (1/\gamma_k)U_k(z^k - \gamma_k U_k^{-1} \nabla g(z^k) - x_f^k) - S x_f^k \in \partial f(x_f^k)
\]

for all \(k \geq 0\).

3. In Algorithm 9, we have

\[
x_g^k := J_{\gamma U^{-1}(\partial g + (1-w)S)}(z^k);
\]

\[
x_f^k := J_{\gamma U^{-1}(\partial f + wS)} \circ \text{refl}_{\gamma U^{-1}(\partial g + (1-w)S)}(z^k);
\]

\[
x_S^k := w x_f^k + (1-w) x_g^k.
\]

\[
\nabla g(x_g^k) := (1/\gamma)U(z^k - x_g^k) - (1-w)S x_g^k \in \partial g(x_g^k);
\]

and \(\nabla f(x_f^k) := (1/\gamma)U(2x_g^k - z^k - x_f^k) - wS x_f^k \in \partial f(x_f^k)\) for all \(k \geq 0\).

4. In Algorithm 10, we have \(\lambda_k = 1\), \(x_g^k := x_f^k, x_S^k := x_f^k\), and

\[
\nabla f(x_f^k) := (1/\gamma_k)U_k(y^k - x_f^k) - S x_f^k \in \partial f(x_f^k)
\]

for all \(k \geq 0\).
Proof. The subgradient identities all follow from Part 1 of Proposition 4.1.1.

Part 1: This is immediate.

Part 2: From Part 1 of Proposition 4.1.1, we have the following identity:

\[ x_f^k = z^k - \gamma_k U_k^{-1} \left( \tilde{\nabla} f(x_f^k) + \nabla g(x_g^k) + Sx_S^k \right). \]

Thus, altogether we have

\[ z^{k+1} = z^k - \gamma_k \lambda_k U_k^{-1} \left( \tilde{\nabla} f(x_f^k) + \nabla g(x_g^k) + Sx_S^k \right). \]

Part 3: We have

\[
\text{refl}_{U^{-1}(\partial f + wS)} \circ \text{refl}_{U^{-1}(\partial g + (1-w)S)}(z^k) \\
= \text{refl}_{U^{-1}(\partial f + wS)}(z^k - 2\gamma U^{-1}(\tilde{\nabla} g(x_g^k) + (1-w)Sx_g^k)) \\
= z^k - 2\gamma U^{-1}(\tilde{\nabla} f(x_f^k) + \tilde{\nabla} g(x_g^k) + S(w x_f^k + (1-w)x_g^k)).
\]

Therefore, if we define \( x_S^k = w x_f^k + (1-w)x_g^k \), then

\[ z^{k+1} = z^k - \gamma \lambda_k U^{-1} \left( \tilde{\nabla} f(x_f^k) + \tilde{\nabla} g(x_g^k) + Sx_S^k \right). \]

Part 4: We have

\[ z^{k+1} - z^k = w^k - y^k = w^k - x_f^k + x_f^k - y^k = -\gamma_k U_k^{-1} \left( \tilde{\nabla} f(x_f^k) + \nabla g(x_f^k) + Sx_f^k \right). \]

\[ \square \]

4.2.2.1 Convergence properties

Now we establish two basic and well known results on the boundedness and summability of various terms related to the above algorithms. These facts will be used repeatedly in our convergence rate analysis.

**Proposition 4.2.1** (Averagedness properties). Let \( U \in S_p(H) \) and let \( \gamma \in \mathbb{R}_{++} \). Then the following hold:

1. The operator \( J_{\gamma U^{-1}(\partial f + S)} \) is \((1/2)\)-averaged in the norm \( \| \cdot \|_U \). In addition, the set of fixed points of \( J_{\gamma U^{-1}(\partial f + S)} \) is equal to \( \text{zer} (\partial f + S) \)
2. Let $\gamma \in (0, 2\beta \rho)$. Suppose that $g$ is differentiable and $\nabla g$ is $(1/\beta)$-Lipschitz. Then the composition

$$T_{FBS}^{U, \gamma} := J_{\gamma U^{-1}(\partial f + S)} \circ (I_H - \gamma U^{-1}\nabla g)$$

(4.2.2)

is $\alpha_{\rho, \gamma}$-averaged in the norm $\| \cdot \|_U$ where

$$\alpha_{\rho, \gamma} := \frac{2\beta \rho}{4\beta \rho - \gamma}. \quad (4.2.3)$$

In addition, the set of fixed points of $T_{FBS}^{U, \gamma}$ is equal to $\text{zer}(\partial f + \nabla g + S)$.

3. Define the PRS operator:

$$T_{PRS} := \text{refl}_{\gamma U^{-1}(\partial f + w S)} \circ \text{refl}_{\gamma U^{-1}(\partial g + (1-w) S)}.$$  \hfill (4.2.4)

Then $T_{PRS}$ is nonexpansive in the metric $\| \cdot \|_U$. Thus, the following DRS operator

$$(T_{PRS})_{1/2} = \frac{1}{2} I_H + \frac{1}{2} \text{refl}_{\gamma U^{-1}(\partial f + w S)} \circ \text{refl}_{\gamma U^{-1}(\partial g + (1-w) S)}$$

(4.2.5)

is $(1/2)$-averaged. In addition, the set of fixed points of $T_{PRS}$ and $(T_{PRS})_{1/2}$ coincide and $\text{zer} (\partial f + \partial g + S) = \{ J_{\gamma U^{-1}(\partial g + (1-w) S)}(z) \mid T_{PRS}z = z \}$.

**Proof.** Parts 1 and 3 are simple modifications of standard facts found in \cite{11}.

**Part 2:** Note that $U^{-1}\nabla g$ is $\beta \rho$-cocoercive by Proposition 4.1.4 and the Baillon-Haddad theorem \cite{3}. Thus, $I_H - \gamma U^{-1}\nabla g$ is $\gamma / (2\beta \rho)$ averaged by \cite{11, Proposition 4.33}. Thus, the formula for $\alpha_{\rho, \gamma}$ follows from \cite[Theorem 3(b)]{98}. The fixed-point identity follows from a simple modification of \cite[Theorem 25.1]{11}. \hfill \Box

**Proposition 4.2.2 (Bounded and summable sequences).** The following hold:

1. Let $z^* \in \text{zer}(\partial f + S)$. Then in Algorithm 7, we have for all $k \geq 0$ that $\|z^{k+1} - z^*\|_{U_{k+1}} \leq (1 + \eta_k)\|z^k - z^*\|^2_{U_k}$ and hence, $\|z^k - z^*\|^2_{U_k} \leq \eta_p\|z^0 - z^*\|^2_{U_0}$.
2. Let \( z^* \in \text{zer}(\partial f + \partial g + S) \). Then in Algorithm 8, the following are true:

   i) For all \( k \geq 0 \), \( \|z^{k+1} - z^*\|_{l_k+1}^2 \leq (1 + \eta_k)\|z^k - z^*\|_{l_k}^2 \) and hence, \( \|z^k - z^*\|_{l_k}^2 \leq \eta_p \|z^0 - z^*\|_{l_0}^2 \).

   ii) The following sum is finite:

\[
\sum_{i=0}^{\infty} \frac{1 - \alpha_i \lambda_i}{\alpha_i \lambda_i} \|z^{i+1} - z^i\|^2 \leq \frac{1}{\rho} (1 + \eta_p \eta_s) \|z^0 - z^*\|_{l_0}^2.
\]

3. Let \( z^* \) be a fixed-point of \( T_{PRS} \). Then in Algorithm 9, we have for all \( k \geq 0 \), that

\( \|z^{k+1} - z^*\|_{l_k+1}^2 \leq \|z^k - z^*\|_{l_k}^2 \) and hence, \( \|z^k - z^*\|_{l_k}^2 \leq \|z^0 - z^*\|_{l_0}^2 \).

4. Let \( z^* \in \text{zer}(\partial f + \partial g + S) \). Then in Algorithm 10, we have for all \( k \geq 0 \) that

\( \|z^{k+1} - z^*\|_{l_k+1}^2 \leq (1 + \eta_k)\|z^k - z^*\|_{l_k}^2 \) and hence, \( \|z^k - z^*\|_{l_k}^2 \leq \eta_p \|z^0 - z^*\|_{l_0}^2 \).

**Proof.** Parts 1, 2, and 3 follow from Proposition 4.1.3 applied to the sequences of operators \( (T_j)_{j \geq 0} := (J_{\gamma_j U_j^{-1}(\partial f + S)})_{j \geq 0} \), \( (T_j)_{j \geq 0} := (T_{FBS}^{\gamma_j})_{j \geq 0} \), and \( (T_j)_{j \geq 0} := ((T_{PRS})_{1/2})_{j \geq 0} \), respectively.

Part 4 follows from Proposition 4.1.5 applied to the the maximal monotone operator \( \partial f \) and the \((\beta^{-1} + \|S\|)\)-Lipschitz operator \( \nabla g + S \).

### 4.2.3 The fundamental inequality

This section describes the pre-primal-dual gap (Definition 4.2.1). We use the pre-primal-dual gap to measure the convergence of the unifying scheme. In Section 4.5, we will show that under certain conditions, the pre-primal-dual gap function bounds the primal and dual objective errors of the iterates generated by a class of primal-dual algorithms.

Before we introduce the gap function, we analyze the optimality conditions of Problem 4.2.1. The following lemma is well-known.

**Lemma 4.2.2.** Let \( x^* \in H \). Suppose that \( x^* \) solves Problem 4.2.1, then for all \( x \in \text{dom}(f) \cap \text{dom}(g) \),

\[
f(x) + g(x) + \langle Sx, -x^* \rangle - f(x^*) - g(x^*) \geq 0.
\]

(4.2.7)
On the other hand, if $\partial (f + g)(x^*) = \partial f(x^*) + \partial g(x^*)$ and $x^*$ satisfies Equation (4.2.7) for all $x \in \text{dom}(f) \cap \text{dom}(g)$, then $x^*$ solves Problem 4.2.1.

Proof. If $x^*$ solves Problem 4.2.1, then $-Sx^*$ is a subgradient of $f + g$ at the point $x^*$. Thus, Equation (4.2.7) follows after noting that $\langle Sx, x \rangle = 0$ for all $x \in H$.

The other direction follows because Equation (4.2.7) characterizes the set of subgradients of the form $-Sx^* \in \partial (f + g)(x^*) = \partial f(x^*) + \partial g(x^*)$. \hfill $\square$

See [11, Corollary 16.38] for conditions that imply additivity of the subdifferential.

Lemma 4.2.2 motivates the following definition:

**Definition 4.2.1 (Pre-primal-dual gap).** Let the setting be as in Algorithm 6. Define the pre-primal dual gap function by the formula: for all $x_f, x_g, x_S, x \in H$, let

$$G^{\text{pre}}(x_f, x_g, x_S; x) = f(x_f) + g(x_g) + \langle Sx_S, -x \rangle - f(x) - g(x).$$

We name $G^{\text{pre}}$ the pre-primal-dual-gap function after the standard primal-dual gap function that appears in [36, 21, 20, 25]. We use the word “pre” because the standard primal-dual gap function usually involves a supremum over the last variable $x$. Note that if $\partial (f + g)(x') = \partial f(x') + \partial g(x')$ and

$$\sup_{x \in H} G^{\text{pre}}(x', x', x'; x) \leq 0,$$

then $x'$ is a solution of Problem 4.2.1 (Lemma 4.2.2).

Our goal throughout the rest of this chapter is to bound the pre-primal-dual gap when $x_f = x_g = x_S$. Because of Equation (4.2.9), all of our upper bounds will be a function of the norm of the last component of $G^{\text{pre}}$. In some cases, we can restrict the supremum in Equation (4.2.9) to a smaller subset $C \subseteq H$. This is the case if, for example, $\text{dom}(f) \cap \text{dom}(g)$ is bounded. Whenever the supremum can be restricted, we obtain a meaningful convergence rate.
Finally, Lemma 4.2.2 shows that for all \( x \in \text{dom}(f) \cap \text{dom}(g) \),

\[
G^\text{pre}(x, x, x; x^*) \geq 0
\]  

whenever \( x^* \) solves Problem 4.2.1. See Section 4.5.1 for other lower bounds of the pre-primal-dual gap in the context of a particular convex optimization problem.

The following is our main tool to bound the pre-primal-dual gap.

**Proposition 4.2.3** (Upper fundamental inequality for primal dual schemes). Suppose that \((z^j)_{j \geq 0}\) is generated by Algorithm 6. Let \( x \in \text{dom}(f) \cap \text{dom}(g) \). Then the following inequality holds: for all \( k \geq 0 \),

\[
2\gamma_k \lambda_k G^\text{pre}(x^*_f, x^*_g; x^*; x) \leq \|z^k - x\|_U^2 - \|z^{k+1} - x\|_U^2 - \|z^{k+1} - z^k\|_U^2 \\
+ 2\gamma_k \lambda_k \langle x^*_f - z^{k+1}, \tilde{\nabla}f(x^*_f) \rangle \\
+ 2\gamma_k \lambda_k \langle x^*_g - z^{k+1}, \tilde{\nabla}g(x^*_g) \rangle \\
+ 2\gamma_k \lambda_k \langle -z^{k+1}, Sx^*_S \rangle.
\]  

(4.2.11)

**Proof.** Fix \( k \geq 0 \). First expand the norm:

\[
\|z^{k+1} - x\|_U^2 = \|z^k - x\|_U^2 + 2\langle x - z^{k+1}, z^{k+1} - z^k \rangle_U - \|z^{k+1} - z^k\|_U^2.
\]

Now we expand the inner product:

\[
2\langle x - z^{k+1}, z^k - z^{k+1} \rangle_U = 2\langle x - z^{k+1}, \gamma_k \lambda_k U^{-1}_k \left( \tilde{\nabla}f(x^*_f) + \tilde{\nabla}g(x^*_g) + Sx^*_S \right) \rangle_U \\
= 2\gamma_k \lambda_k \langle x - z^{k+1}, \tilde{\nabla}f(x^*_f) \rangle + 2\gamma_k \lambda_k \langle x - z^{k+1}, \tilde{\nabla}g(x^*_g) \rangle \\
+ 2\gamma_k \lambda_k \langle x - z^{k+1}, Sx^*_S \rangle.
\]

We add and subtract a point in the inner products involving \( f \) and \( g \) and use the subgradient inequality to get:

\[
2\gamma_k \lambda_k \langle x - z^{k+1}, \tilde{\nabla}f(x^*_f) \rangle \leq 2\gamma_k \lambda_k \langle x^*_f - z^{k+1}, \tilde{\nabla}f(x^*_f) \rangle + 2\gamma_k \lambda_k (f(x) - f(x^*_f)); \\
2\gamma_k \lambda_k \langle x - z^{k+1}, \tilde{\nabla}g(x^*_g) \rangle \leq 2\gamma_k \lambda_k \langle x^*_g - z^{k+1}, \tilde{\nabla}g(x^*_g) \rangle + 2\gamma_k \lambda_k (g(x) - g(x^*_g)).
\]

Therefore Equation (4.2.11) follows after rearranging. \( \square \)
The upper fundamental inequality in Proposition 4.2.3 bounds the pre-primal-dual gap with the sum of an alternating sequence and a key term.

**Definition 4.2.2 (Upper key term).** Let \((z^j)_{j\geq 0}\) be generated by Algorithm 6. For all \(k \geq 0\), we define the fundamental upper key term

\[
\kappa_u^k(\lambda_k) := -\|z^{k+1} - z^k\|_{U_k}^2 + 2\gamma_k \lambda_k \langle x^k_f - z^{k+1}, \nabla f(x^k_f) \rangle + 2\gamma_k \lambda_k \langle x^k_g - z^{k+1}, \nabla g(x^k_g) \rangle + 2\gamma_k \lambda_k \langle -z^{k+1}, S x^k_S \rangle. \tag{4.2.12}
\]

The value \(\kappa_u^k(\lambda_k)\) depends on the entire history of Algorithm 6 up to and including iteration \(k\), but in our analysis we will only view \(\kappa_u^k(\lambda_k)\) as a function of the parameter \(\lambda_k\). Throughout the rest of the chapter, we will often make the dependence of the upper key term on \(\lambda_k\) implicit, and denote \(\kappa_u^k := \kappa_u^k(\lambda_k)\). However, in the proof of Theorem 4.4 we will need to keep the dependence explicit.

### 4.2.3.1 Computing the upper key terms

The following proposition will compute the upper key terms induced by the PPA, FBS, PRS, and FBF algorithms. See Section 4.2.2 for the definitions of the points \(x^k_f, x^k_g,\) and \(x^k_S\).

**Proposition 4.2.4 (Computing the upper key terms).** Let \((z^j)_{j\geq 0}\) be generated by Algorithm 6. Then the following simplifications of the upper key terms can be made

1. In Algorithm 7, we have \(\kappa_u^k(\lambda_k) = (1 - 2/\lambda_k) \|z^{k+1} - z^k\|_{U_k}^2\).

2. In Algorithm 8, we have

\[
\kappa_u^k(\lambda_k) \leq \left( \rho - \frac{\varepsilon}{\beta \lambda_k} \right) \|z^{k+1} - z^k\|^2 + 2\gamma_k \lambda_k g(x^k_g) - 2\gamma_k \lambda_k g(x^k_f).
\]

3. In Algorithm 9, we have \(\kappa_u^k(\lambda_k) = (1 - 2/\lambda_k) \|z^{k+1} - z^k\|_{U_k}^2\).
4. In Algorithm 10, we have \( \kappa^k_u(\lambda_k) \leq 0 \).

Proof. To simplify notation, we drop the iteration index and denote \( z = z^k, x_f := x^k_f, x_g := x^k_g, x_S := x^k_S, z^+ := z^{k+1}, \gamma := \gamma_k, \lambda := \lambda_k, U := U_k, \) and \( \kappa_u := \kappa^k_u(\lambda_k) \) throughout this proof.

For PPA, FBS, and PRS, we note that the following identities hold:

\[
z^+ - z = \lambda (x_f - x_g),
\]
and there exists \( w \in \mathbb{R} \) such that

\[
x_S = w x_f + (1 - w) x_g.
\]

Indeed, in PPA and FBS, \( w = 1 \) (see Section 4.2.2). In PRS, \( w \) is a parameter of the algorithm, and Equations (4.2.14) and (4.2.13) are shown in Lemma 4.2.1. Furthermore, Part 1 of Proposition 4.1.1 shows that in PPA and FBS,

\[
x_f = x_g - \gamma U^{-1} \left( \nabla f(x_f) + \nabla g(x_g) + Sx_S \right)
\]
for a unique subgradient \( \nabla f(x_f) \in \partial f(x_f) \); see Lemma 4.2.1 for the definition of \( \nabla f(x_f) \).

Now we claim that in PPA, FBS, and PRS,

\[
\kappa_u = 2 \langle x_f + \gamma U^{-1} (\nabla g(x_g) + (1 - w) Sx_g) - z^+, z - z^+ \rangle_U - \|z^+ - z\|_U^2
\]
where we make the identification \( \nabla g(x_g) = \nabla g(x_g) \) whenever \( g \) is differentiable; see Lemma 4.2.1 for the definition of \( \nabla g(x_g) \). Because \( x_S = x_g + w(x_f - x_g) = x_g + (w/\lambda)(z^+ - z) \) and \( \langle Sx, x \rangle = 0 \) for all \( x \in H \), we have the simplification:

\[
2 \langle z - z^+, \gamma(1 - w) Sx_S \rangle = 2 \langle z - z^+, \gamma(1 - w) Sx_g \rangle.
\]
Therefore,

\[
\kappa_u = -\|z^+ - z\|_U^2 + 2\gamma\lambda \langle x_f - z^+, \nabla f(x_f) \rangle \\
+ 2\gamma\lambda \langle x_g - z^+, \nabla g(x_g) \rangle + 2\gamma\lambda \langle x_s - z^+, Sx_s \rangle \\
= -\|z^+ - z\|_U^2 + 2\gamma\lambda \langle x_f - z^+, \nabla f(x_f) \rangle \\
+ 2\gamma\lambda \langle x_g - x_f, \nabla g(x_g) \rangle + 2\gamma\lambda \langle x_f - z^+, \nabla g(x_g) \rangle \\
+ 2\gamma\lambda \langle x_s - x_f, Sx_s \rangle + 2\gamma\lambda \langle x_f - z^+, Sx_s \rangle \\
= -\|z^+ - z\|_U^2 + 2\gamma\lambda \langle x_f - z^+, \nabla f(x_f) + \nabla g(x_g) + Sx_s \rangle \\
= 2\langle x_f - z^+, z - z^+ \rangle_U \\
+ 2\langle z - z^+, \gamma \nabla g(x_g) \rangle + 2\langle z - z^+, \gamma(1 - w)Sx_s \rangle \\
= 2\langle x_f + \gamma U^{-1}(\nabla g(x_g) + (1 - w)Sx_s) - z^+, z - z^+ \rangle_U - \|z^+ - z\|_U^2
\]

where the second to last equality uses Equation (4.2.15) and the second to last “+” uses Equation (4.2.13).

Now we proceed with the specific cases: In PPA and FBS, \( w = 1 \) and

\[
\kappa_u \overset{(4.2.16)}{=} 2\langle x_f + \gamma U^{-1}\nabla g(x_g) - z^+, z - z^+ \rangle_U - \|z^+ - z\|_U^2 \\
= 2\langle x_f - z^+, z - z^+ \rangle_U + 2\gamma \langle \nabla g(x_g), z - z^+ \rangle - \|z^+ - z\|_U^2 \\
= 2 \left(1 - \frac{1}{\lambda}\right) \|z^+ - z\|_U^2 + 2\gamma\lambda \langle \nabla g(x_g), x_g - x_f \rangle - \|z^+ - z\|_U^2 \\
\leq \left(1 - \frac{2}{\lambda}\right) \|z^+ - z\|_U^2 + 2\gamma\lambda g(x_g) - 2\gamma\lambda g(x_f) + \frac{\gamma}{\lambda\beta} \|z^+ - z\|^2
\]

where the third inequality follows from \( x_f - z^+ = (1 - (1/\lambda))(z - z^+) \), we use the identity \( z^+ - z = \lambda(x_f - x_g) \) (Equation (4.2.13)) on the last two lines, and the last inequality follows from the Descent Theorem [11, Theorem 18.15(iii)]: \( \langle \nabla g(x_g), x_g - x_f \rangle \leq g(x_g) - g(x_f) + (1/(2\beta))\|x_g - x_f\|^2 \). In PPA \( g \equiv 0 \), so the Equation (4.2.18) implies the identity in Part 1. The inequality for FBS now follows by the above bound for \( \kappa_u \), the bound
\[ \gamma \leq 2\beta \rho - \varepsilon, \quad \text{and} \]
\[
\left(1 - \frac{2}{\lambda}\right) \|z^+ - z\|_U^2 + \frac{\gamma}{\lambda \beta} \|z^+ - z\|^2 \leq \left(\rho + \frac{\gamma - 2\beta \rho}{\lambda \beta}\right) \|z^+ - z\|^2
\]

where we use \(\lambda \leq (4\beta \rho - \gamma)/2\beta \rho \leq 2\) and the lower bound \(U \succcurlyeq \rho I_{\mathbf{H}}\).

For relaxed PRS, we have
\[
z^+ = z + \lambda (x_f - x_g) = (1 - \lambda)z + \lambda (x_f - x_g + z)
\]
\[
= (1 - \lambda)z + \lambda \left(x_f + \gamma U^{-1} \left(\tilde{\nabla}g(x_g) + (1 - w)Sx_g\right)\right).
\]

Therefore, subtract \(\lambda z^+ + (1 - \lambda)z\) from both sides of the above equation, divide by \(\lambda\), and use the identity in Equation (4.2.16) to get
\[
\kappa_u = 2\langle x_f + \gamma U^{-1} \left(\tilde{\nabla}g(x_g) + (1 - w)Sx_g\right) - z^+, z - z^+\rangle_U - \|z^+ - z\|_U^2
\]
\[
= 2 \left(1 - \frac{1}{\lambda}\right) \|z^+ - z\|_U^2 - \|z^+ - z\|_U^2 = \left(1 - \frac{2}{\lambda}\right) \|z^+ - z\|_U^2.
\]

Finally, we prove the bound for the FBF algorithm:
\[
\kappa_u = 2\langle x_f - x_f, \gamma \tilde{\nabla}f(x_f) + \gamma \nabla g(x_f) + \gamma Sx_f\rangle - \|z^+ - z\|_U^2
\]
\[
= 2\langle x_f - z^+, z - z^+\rangle_U - \|z^+ - z\|_U^2 \quad \text{(4.1.7)} \|x_f - z^+\|_U^2 - \|x_f - z\|_U^2.
\]

Furthermore, the identity holds:
\[
z^+ - x_f = z^+ - z + z - x_f
\]
\[
= -\gamma U^{-1} \left(\tilde{\nabla}f(x_f) + \nabla g(x_f) + Sx_f\right) + \gamma U^{-1} \left(\tilde{\nabla}f(x_f) + \nabla g(z) + Sz\right)
\]
\[
= \gamma U^{-1} (\nabla g(z) + Sz - \nabla g(x_f) - Sx_f).
\]

Note that the operator \(\nabla g + S\) is \((1/\beta) + \|S\|\) Lipschitz. Thus,
\[
\|x_f - z^+\|_U^2 - \|x_f - z\|_U^2 \quad \text{(4.2.19)} \quad \gamma^2 \|\nabla g(z) + Sz - \nabla g(x_f) - Sx_f\|_{U^{-1}}^2 - \|x_f - z\|_U^2
\]
\[
\leq \left(\frac{\gamma^2}{\rho} \left(\frac{1}{\beta} + \|S\|\right)^2 - \rho\right) \|x_f - z\|^2 \leq 0,
\]

where we use the following bound: for all \(x \in \mathbf{H}, \|x\|_{U^{-1}}^2 \leq (1/\rho)\|x\|^2\) (Lemma 4.1.1). \(\square\)
4.3 Ergodic convergence

In this section, we prove an ergodic convergence rate for the pre-primal-dual gap. To this end, we recall the partial sum sequence $\Sigma_k = \sum_{i=0}^{k} \gamma_i \lambda_i$, and for every sequence of vectors $(x^i)_{j \geq 0} \subseteq H$, we define the ergodic sequence $x^k = \frac{1}{\Sigma_k} \sum_{i=0}^{k} \gamma_i \lambda_i x^i$. For each algorithm, Theorem 4.3.1 (below) gives an ergodic sequence $(x^j)_{j \geq 0}$ such that for all bounded subsets $D \subseteq H$, we have

$$
\sup_{x \in D} G_{pre}(x^k, x^k, x^k ; x) = O\left( \frac{1 + \sup_{x \in D} \|x\|^2}{\Sigma_k} \right).
$$

This bound is a generalization of the primal-dual gap bounds shown in [36, 21, 20, 25]. See Section 4.5.1 for several lower bounds of the pre-primal-dual gap.

Before we prove our ergodic rates, we need to prove a bound for PRS. Recall that we only analyze the PRS algorithm when the map $U_k \equiv U$ is fixed. The following lemma will help us deduce the convergence rate of the PRS algorithm whenever $f$ or $g$ is Lipschitz (Part 3 of Theorem 4.3.1).

**Lemma 4.3.1.** Suppose that $(z^j)_{j \geq 0}$ is generated by the relaxed PRS algorithm and that $z^*$ is a fixed-point of $T_{PRS}$ (see equation (4.2.4)). Then the following ergodic bound holds:

$$
\|x^k_f - x^k_g\|_U \leq \frac{2\|z^0 - z^*\|_U}{\Sigma_k}.
$$

**Proof.** The identity $\lambda_k(x^k_f - x^k_g) = z^{k+1} - z^k$ and the fact the sequence $(\|z^j - z^*\|_U)_{j \geq 0}$ is decreasing (Part 3 of Proposition 4.2.2), show that

$$
\|x^k_f - x^k_g\|_U = \left\| \frac{\gamma}{\Sigma_k} \sum_{i=0}^{k} \lambda_i (x^i_f - x^i_g) \right\|_U = \frac{\gamma \|z^{k+1} - z^0\|_U}{\Sigma_k} \leq \frac{\gamma \|z^{k+1} - z^*\|_U + \gamma \|z^0 - z^*\|_U}{\Sigma_k} \leq \frac{2\|z^0 - z^*\|_U}{\Sigma_k}.
$$

Lemma 4.3.1 shows that the difference of splitting variables $x^k_f - x^k_g$ converges to zero with rate $O(1/\Sigma_k)$. Thus, if $f$ is Lipschitz continuous, then $|f(x^k_f) - f(x^k_g)| = O(1/\Sigma_k)$.

We are now ready to prove our main ergodic convergence results.
Theorem 4.3.1 (Ergodic convergence of the unifying primal dual scheme). Suppose that the sequence \((z_j)_{j \geq 0}\) is generated by Algorithm 6, and suppose that Assumption 4.1.3 holds. Then for all \(x \in H\) and all \(k \geq 0\), we have the following:

1. **Ergodic convergence of PPA:** Let \(z^* \in \text{zer}(\partial f + S)\). Then in Algorithm 7, we have
   \[
   G^{\text{pre}}(x_f, 0, x_f^k; x) \leq \frac{\|z^0 - x\|_{U_0}^2 + 2\eta_p \eta_h \|z^0 - z^*\|_{U_0}^2 + 2\mu \eta_h \|z^* - x\|^2}{2\Sigma_k}.
   \]

2. **Ergodic convergence of FBS:** Let \(z^* \in \text{zer}(\partial f + \partial g + S)\), and let \(\bar{\lambda} = \sup_{j \geq 0} \lambda_j\). Then in Algorithm 8, we have
   \[
   G^{\text{pre}}(x_f, x_f^k, x_f^k; x) \leq \frac{\left(\|z^0 - x\|_{U_0}^2 + \left(2\eta_p \eta_h + \frac{(1+\eta_p \eta_h) \max \{\rho - \epsilon/(\beta \lambda), 0\}}{\rho \inf_{j \geq 0} (1 - \alpha_j \lambda_j)/(\alpha_j \lambda_j)}\right) \|z^0 - z^*\|_{U_0}^2 + 2\mu \eta_h \|z^* - x\|^2\right)}{2\Sigma_k}.
   \]

3. **Ergodic convergence of PRS:** Let \(z^*\) be a fixed point of \(T_{PRS}\). Suppose that \(f\) (respectively \(g\)) is \(L\)-Lipschitz, let \(x^k := x^k_f\) (respectively \(x^k := x^k_f\)), and let \(\hat{w} = w\) (respectively \(\hat{w} = 1 - w\)). Then in Algorithm 9, we have
   \[
   G^{\text{pre}}(x_f^k, x_f^k, x_f^k; x) \leq \frac{\|z^0 - x\|_{U_0} + 4(\gamma/\rho)(L + |\hat{w}|\|S\|\|x\|)\|z^0 - z^*\|_{U_0}}{2\Sigma_k}.
   \]

4. **Ergodic convergence of FBF:** Let \(z^* \in \text{zer}(\partial f + \partial g + S)\). Then in Algorithm 10, we have
   \[
   G^{\text{pre}}(x_f, x_f^k, x_f^k; x) \leq \frac{\|z^0 - x\|_{U_0}^2 + 2\eta_p \eta_h \|z^0 - z^*\|_{U_0}^2 + 2\mu \eta_h \|z^* - x\|^2}{2\Sigma_k}.
   \]

Proof. For any sequence of points \((z^j)_{j \geq 0} \subseteq H\) and any point \(z^* \in H\) such that \(\|z^{k+1} - z^*\|_{U_{k+1}} \leq (1+\eta_k)\|z^k - z^*\|_{U_k}^2\) for all \(k \geq 0\), we have \(\|z^k - z^*\|_{U_k}^2 \leq (\prod_{i=0}^{\infty} (1 + \eta_i)) \|z^0 - z^*\|_{U_0}^2\).

Therefore, by the convexity of \(\|\cdot\|_{U_i}\) for all \(i \geq 0\), and by the inequality \(-\|x\|_{U_i} \leq \frac{\|z^0 - x\|_{U_0}^2 + 2\eta_p \eta_h \|z^0 - z^*\|_{U_0}^2 + 2\mu \eta_h \|z^* - x\|^2}{2\Sigma_k}\).
\[-(1/(1 + \eta_i))\|x\|_{U_i} for all x \in H and i \geq 0, we have for all k \geq 0,
\[
\sum_{i=0}^{k} (\|z^i - x\|_{U_i}^2 - \|z^{i+1} - x\|_{U_i}^2)
\leq \|z^0 - x\|_{U_0}^2 + \sum_{i=0}^{k} (\|z^{i+1} - x\|_{U_{i+1}}^2 - \|z^{i+1} - x\|_{U_i}^2)
\leq \|z^0 - x\|_{U_0}^2 + \sum_{i=0}^{k} \frac{\eta_i}{1 + \eta_i} \|z^{i+1} - x\|_{U_{i+1}}^2
\leq \|z^0 - x\|_{U_0}^2 + \sum_{i=0}^{k} \eta_i \left(\|z^{i+1} - z^*\|_{U_{i+1}}^2 + \|z^* - x\|_{U_{i+1}}^2\right)
\leq \|z^0 - x\|_{U_0}^2 + \left(2 \left(\prod_{i=0}^{\infty} (1 + \eta_i)\right) \sum_{i=0}^{\infty} \eta_i \|z^0 - z^*\|_{U_0}^2 + 2\mu \left(\sum_{i=0}^{\infty} \eta_i \right) \|z^* - x\|^2\right) = \|z^0 - x\|_{U_0}^2 + 2\eta_p \eta_s \|z^0 - z^*\|_{U_0}^2 + 2\mu \eta_s \|z^* - x\|^2.
\]

We will use Equation (4.3.2) to produce bounds for all of the variable metric methods.

Part 1: This follows from the Jensen’s inequality, Proposition 4.2.4 (\(\kappa_u^i = (1 - 2/\lambda_i) \|z^{i+1} - z^i\|_{U_i} \leq 0\)), and the fundamental inequality:

\[
\mathcal{G}^{\text{pre}}(x_i^k, 0, x_i^k; x) \leq \frac{1}{\Sigma_k} \sum_{i=0}^{k} \gamma_i \lambda_i \mathcal{G}^{\text{pre}}(x_i, 0, x_i^i; x)
\leq \frac{1}{2\Sigma_k} \sum_{i=0}^{k} (\kappa_u^i + \|z^i - x\|_{U_i}^2 - \|z^{i+1} - x\|_{U_i}^2)
\leq \frac{1}{2\Sigma_k} \left(\|z^0 - x\|_{U_0}^2 + 2\eta_p \eta_s \|z^0 - z^*\|_{U_0}^2 + 2\mu \eta_s \|z^* - x\|^2\right).
\]

Part 2: We have the following bound from Proposition 4.2.2:

\[
\sum_{i=0}^{k} \left(\rho - \frac{\varepsilon}{\beta \lambda_i}\right) \|z^{i+1} - z^i\|^2 \leq \frac{\max \{\rho - \varepsilon/(\beta \lambda_i)\}}{\rho \inf_{j \geq 0} (1 - \alpha_j \lambda_j)/(\alpha_j \lambda_j)} \left(1 + \eta_p \eta_s\right) \|z^0 - z^*\|_{U_0}^2. 
\]

Thus, the bound follows from Jensen’s inequality, Proposition 4.2.4

\((\kappa_u^i = (\rho - \varepsilon/(\beta \lambda_i)) \|z^{i+1} - z^i\|^2 + 2\gamma_i \lambda_i g(x_i^g) - 2\gamma_i \lambda_i g(x_i^g)>>), and the fundamental inequl-
ity:

\[ G^\text{pre}(\bar{x}_f^k, \bar{x}_g^k, \bar{x}_f^k; x) \leq \frac{1}{\sum_k} \sum_{i=0}^k \gamma_i \lambda_i G^\text{pre}(x_f^i, x_g^i, x_f^i; x) \]

\[ = \frac{1}{\sum_k} \sum_{i=0}^k (\gamma_i \lambda_i G^\text{pre}(x_f^i, x_g^i, x_f^i; x) + \gamma_i \lambda_i g(x_f^i) - \gamma_i \lambda_i g(x_g^i)) \]

\[ (4.2.11) \]

\[ \leq \frac{1}{2\sum_k} \sum_{i=0}^k \left( \kappa_u^i + \|z^i - x\|_{U_0}^2 - \|z^{i+1} - x\|_{U_0}^2 + 2\gamma_i \lambda_i g(x_f^i) - 2\gamma_i \lambda_i g(x_g^i) \right) \]

\[ \leq \frac{1}{2\sum_k} \left( \sum_{i=0}^k \left( \rho - \frac{\varepsilon}{\beta \lambda_i} \right) \|z^{i+1} - z_i\|^2 \right) \]

\[ + \frac{1}{2\sum_k} \left( \|z^0 - x\|_{U_0}^2 + 2\eta_p \eta_h \|z^0 - z^*\|_{U_0}^2 + 2\mu \eta_h \|z^* - x\|^2 \right) \]

\[ \leq \left( \|z^0 - x\|_{U_0}^2 + \left( 2\eta_p \eta_h + \frac{(1+\eta_p \eta_h) \max \{ \rho - \varepsilon/\left( \beta \lambda_i \right) \} }{\rho \inf_{j \geq 0} (1-\alpha_j \lambda_j)/(\alpha_j \lambda_j)} \right) \|z^0 - z^*\|_{U_0}^2 + 2\mu \eta_h \|z^* - x\|^2 \right) \]

\[ \frac{2\sum_k}{(4.3.3)} \]

Part 3: We prove the result when \( f \) is Lipschitz; the other case is symmetric. This follows from the Jensen’s inequality, Proposition 4.2.4

\[ (\kappa_u^i = (1-2/\lambda_i) \|z^{i+1} - z_i\|_U^2 \leq 0) \], the fundamental inequality, and the identity \( \bar{x}_g^k - \bar{x}_S^k = w(\bar{x}_g^k - \bar{x}_f^k) \):

\[ G^\text{pre}(\bar{x}_g^k, \bar{x}_g^k, \bar{x}_g^k; x) = G^\text{pre}(\bar{x}_f^k, \bar{x}_g^k, \bar{x}_S^k; x) + f(\bar{x}_g^k) - f(\bar{x}_f^k) + \langle S(\bar{x}_g^k - \bar{x}_S^k), -x \rangle \]

\[ \leq \frac{1}{\sum_k} \sum_{i=0}^k \gamma_i \lambda_i G^\text{pre}(x_f^i, x_g^i, x_f^i; x) \]

\[ + f(\bar{x}_g^k) - f(\bar{x}_f^k) + \langle S(\bar{x}_g^k - \bar{x}_S^k), -x \rangle \]

\[ (4.2.11) \]

\[ \leq \frac{1}{2\sum_k} \sum_{i=0}^k \left( \kappa_u^i + \|z^i - x\|_{U_0}^2 - \|z^{i+1} - x\|_{U_0}^2 \right) \]

\[ + L \|\bar{x}_g^k - \bar{x}_f^k\| + \|S\| \|\bar{x}_g^k - \bar{x}_S^k\| \|x\| \]

\[ \leq \left( \|z^0 - x\|_{U_0}^2 + 4(\gamma/\rho) (L + \|w\| S \|\|x\|\|) \|z^0 - z^*\|_U \right) \]

\[ \frac{2\sum_k}{(4.3.1)} \]

Part 4: This follows from the Jensen’s inequality, Proposition 4.2.4 (\( \kappa_u^i \leq 0 \)), and the
fundamental inequality:

\[
G_{\text{pre}}(x^k_f, x^k_f, x^k_f; x) \leq \frac{1}{\sum_k} \sum_{i=0}^k \gamma_i \lambda_i G_{\text{pre}}(x^i_f, x^i_f, x^i_f; x)
\]

\[\leq \frac{1}{2\sum_k} \sum_{i=0}^k \left( \kappa^i_u + \| z^i - x \|_{U_i}^2 - \| z^{i+1} - x \|_{U_i}^2 \right) \]

\[\leq \frac{1}{2\sum_k} \left( \| z^0 - x \|_{U_0}^2 + 2\eta_p \eta_s \| z^0 - z^* \|_{U_0}^2 \right) + 2\mu \eta_s \| z^* - x \|_2^2 \right). \]

\[\square\]

**Remark 4.3.1.** The PPA, FBS, and PRS, rates in Theorem 4.3.1 are sharp (up to constants). See Proposition 2.7.2 of Chapter 2 for the result when \( S \) is the zero map.

### 4.4 Nonergodic convergence

In this section we deduce nonergodic convergence rates for PPA, FBS and PRS under the following assumption:

**Assumption 4.4.1.** For all nonergodic convergence results, we fix the metrics and the implicit stepsize parameters.

For PPA, FBS, and PRS, Theorem 4.4.2 (below) produces a natural sequence \((x^j)_{j\geq0}\) such that for all bounded subsets \(D \subseteq H\), we have

\[
\sup_{x \in D} G_{\text{pre}}(x^k, x^k, x^k; x) = o \left( \frac{1 + \sup_{x \in D} \| x \|_U}{\sqrt{k+1}} \right).
\]

To the best of our knowledge, the rate of convergence for the nonergodic primal-dual gap generated by the class of algorithms we study has never appeared in the literature.

Nonergodic iterates tend to share structural properties, such as sparsity or low rank, with the solution of the problem. In some cases, the ergodic iterates generated in Section 4.3 “average out” structural properties of the nonergodic iterates. Thus, although the ergodic iterates may be “closer” to the solution, they are often poorer partial solutions than the nonergodic iterates. The results of this section provide worst-case theoretical
guarantees on the quality of the nonergodic iterates in order to justify their use in practical applications.

In our analysis, we use the following result:

**Theorem 4.4.1** (Theorem 2.3.1 of Chapter 2). Let $\alpha \in (0, 1)$ and let $U \in S_\rho(H)$. Suppose that $T : H \to H$ is an $\alpha$-averaged operator in the norm $\| \cdot \|_U$. Let $z^*$ be a fixed point of $T$, let $z^0 \in H$, let $\tau_k := (1 - \alpha \lambda_k) \lambda_k / \alpha$ for all $k \geq 0$, suppose that $\tau := \inf_{j \geq 0} \tau_j > 0$, and suppose that $(z^j)_{j \geq 0}$ is generated by the following iteration:

$$z^{k+1} := T_{\lambda_k}(z^k). \quad (4.4.1)$$

Then

$$\|Tz^k - z^k\|_U^2 \leq \|z^0 - z^*\|_U^2 / \tau(k + 1) \quad \text{and} \quad \|Tz^k - z^k\|_U^2 = o \left( \frac{1}{k + 1} \right). \quad (4.4.2)$$

Throughout this section, the letter $T$ will always denote an $\alpha$-averaged mapping in the norm $\| \cdot \|_U$. Recall that for all $\lambda \in (0, 1/\alpha]$, $T_\lambda$ is $\alpha \lambda$ averaged (see Proposition 4.1.1), so

$$\|T_\lambda z^k - z^*\|_U^2 \leq \|z^k - z^*\|_U^2 - \frac{1 - \alpha \lambda}{\alpha \lambda} \|T_\lambda z^k - z^k\|_U^2 \quad (4.4.3)$$

for any fixed-point $z^*$ of $T$. Equation (4.4.3) shows that $T_\lambda z^k$ is at least as close to $z^*$ as $z^k$ is. This fact will be useful in the proof of Theorem 4.4.2 below.

In the following theorem, we will deduce little-$o$ and big-$O$ convergence rates. Because the pre-primal-dual gap can be negative, we slightly abuse notation: a (not necessarily positive) sequence $(a_j)_{j \geq 0}$ satisfies $a_k = o((1 + \|x\|_U) / \sqrt{k + 1})$ provided that there exists a nonnegative sequence $(b_j)_{j \geq 0}$ such that $b_k = o((1 + \|x\|_U) / \sqrt{k + 1})$ and $a_k = O(b_k)$. Note that we do not measure $|a_k|$ because our only goal is to ensure that the sequence $(a_j)_{j \geq 0}$ is eventually nonpositive.

**Theorem 4.4.2.** Suppose that Assumption 4.4.1 holds, let $U \in S_\rho(H)$ denote the common metric inducing map, and let $\gamma \in R_{++}$ denote the common stepsize parameter. Then each method is a special case of Iteration (4.4.1). For each method, assume that $\tau > 0$ (See Theorem 4.4.1). Then for all $k \geq 0$ and all $x \in H$, the following hold:
1. **Nonergodic convergence of PPA:** Let \( z^* \in \text{zer}(\partial f + S) \). Then in Algorithm 7, we have \( \alpha = 1/2 \) and \( T = J_{U^{-1}(\partial f + S)} \),

\[
G^\text{pre}(x^k_T, 0, x^k_G; x) \leq \frac{\|z^0 - z^*\|_U + \|z^* - x\|_U \|z^0 - z^*\|_U}{\gamma \sqrt{\tau}(k+1)},
\]

and \( G^\text{pre}(x^k_T, 0, x^k_G; x) = o\left((1 + \|x\|_U)/\sqrt{k+1}\right) \).

2. **Nonergodic convergence of FBS:** Let \( z^* \in \text{zer}(\partial f + \partial g + S) \). Then in Algorithm 8, we have \( \alpha = \alpha_{\gamma, \rho} \) (Equation (4.2.3)) and \( T = T_{U, \gamma}^{FBS} \) (Equation (4.2.2)),

\[
G^\text{pre}(x^k_T, x^k_F, x^k_G; x) \leq \frac{\|z^0 - z^*\|_U + \|z^* - x\|_U \|z^0 - z^*\|_U}{\gamma \sqrt{\tau}(k+1)},
\]

and \( G^\text{pre}(x^k_T, x^k_F, x^k_G; x) = o\left((1 + \|x\|_U)/\sqrt{k+1}\right) \).

3. **Nonergodic convergence of PRS:** Let \( z^* \) be a fixed point of \( T_{PRS} \) (Equation (4.2.4)). Then in Algorithm 9, we have \( \alpha = 1/2 \) and \( T = (T_{PRS})_{1/2} \) (Equation (4.2.5)). In addition, suppose that \( f \) (respectively \( g \)) is \( L \)-Lipschitz, let \( x^k := x^k_g \) (respectively \( x^k := x^k_f \)), and let \( \hat{w} = w \) (respectively \( \hat{w} = 1 - w \)). Then

\[
G^\text{pre}(x^k, x^k, x^k; x) \leq \frac{\|z^0 - z^*\|_U + \|z^* - x\|_U + (\gamma/\rho)(L + \|\hat{w}\|_S\|x\|_S) \|z^0 - z^*\|_U}{\gamma \sqrt{\tau}(k+1)},
\]

and \( G^\text{pre}(x^k, x^k, x^k; x) = o\left((1 + \|x\|_U)/\sqrt{k+1}\right) \).

**Proof.** Fix \( k \geq 0 \). In all of the following proofs, we will bound the pre-primal-dual gap by a quantity involving \( \|Tz^k - z^k\|_U \). Then the big-\( O \) and little-\( o \) convergence rates follow directly from Theorem 4.4.1. In addition, we will use Equation (4.4.3) and the independence of \( x^k_T, x^k_G, x^k_S \) from \( \lambda_k \) to tighten our upper bounds. To this end, we will denote \( z^*_\lambda := T_\lambda(z^k) \) (see Equation (4.1.3)) and let \( C = (0, 1/\alpha) \) where \( \alpha \) is averagedness coefficient of \( T \). Note that \( T_\lambda \) is nonexpansive for all \( \lambda \in C \) (see Proposition (3)). Also note that for \( \lambda \in C \), we have \((1/\lambda)(z^*_\lambda - z^*) = Tz^k - z^k \) and \( \|z^*_\lambda - z^*\|_U \leq \|z^k - z^*\|_U \leq \|z^0 - z^*\|_U \) by Equation (4.4.3) and the monotonicity of \( \|z^j - z^*\|_U \) \( j \geq 0 \) (Proposition 4.2.2). Thus,
\[ \|z_\lambda - x\|_U \leq \|z^0 - z^*\|_U + \|z^* - x\|_U. \] Therefore, for all \( \lambda \in (0, 1/\alpha] \), we have

\[
\frac{\langle z^k - z_\lambda, z_\lambda - x \rangle_U}{\lambda} \leq \|Tz^k - z^k\|_U \|z_\lambda - x\|_U \leq \frac{\|z^0 - z^*\|_U + \|z^* - x\|_U}{\sqrt{\tau(k + 1)}}. \tag{4.4.4}
\]

Note that the upper key term identities (Proposition 4.2.4) and the fundamental inequality (Proposition 4.2.3) continue to hold when \( z^{k+1} \) is replaced by \( z_\lambda \). Thus, in each of the cases below, we will minimize the fundamental inequality over all \( \lambda \in C \).

**Part 1:** Proposition 4.2.4 shows that \( \kappa_\mu^k(\lambda) = (1 - 2/\lambda) \|z_\lambda - z^k\|^2_U \). Thus, the fundamental inequality, the cosine rule, and the identity \( C = (0, 2] \) show

\[
G^{pre}(x^k, 0, x^k; x) \leq \inf_{\lambda \in C} \frac{1}{2\gamma \lambda} \left( 1 - \frac{2}{\lambda} \right) \|z_\lambda - z^k\|^2_U + \|z^k - x\|^2_U - \|z_\lambda - x\|^2_U \tag{4.1.7}
\]

\[
\leq \frac{1}{\gamma} \langle z^k - z_1, z_1 - x \rangle_U \leq \frac{\|z^0 - z^*\|_U + \|z^* - x\|_U}{\gamma \sqrt{\tau(k + 1)}} \frac{\|z^0 - z^*\|_U}{\gamma \sqrt{\tau(k + 1)}}. \tag{4.4.4}
\]

**Part 2:** First choose \( \tilde{\lambda} \in C \) small enough that \((\rho + \mu) - \varepsilon / (\beta \tilde{\lambda}) \leq 0\). Now recall that Proposition 4.2.4 proves the following inequality: \( \kappa_\mu^k(\lambda) \leq (\rho - \varepsilon / (\beta \lambda)) \|z^{k+1} - z^k\|^2 + 2\gamma \lambda g(x^k) - 2\gamma \lambda g(x^k) \). Thus, the fundamental inequality, the cosine rule, and the identity \( C = (0, 1/\alpha] \) show

\[
G^{pre}(x^k, x^k, x^k; x) = G^{pre}(x^k, x^k, x^k; x) + g(x^k) - g(x^k)
\]

\[
\leq \inf_{\lambda \in C} \frac{1}{2\gamma \lambda} \left( 2\gamma \lambda g(x^k) - 2\gamma \lambda g(x^k) + \kappa_\mu^k(\lambda) + \|z^k - x\|^2_U - \|z_\lambda - x\|^2_U \right)
\]

\[
\leq \inf_{\lambda \in C} \frac{1}{2\gamma \lambda} \left( \left( \frac{\rho - \varepsilon}{\beta \lambda} \right) \|z_\lambda - z^k\|^2 + \|z^k - x\|^2_U - \|z_\lambda - x\|^2_U \right) \tag{4.1.7}
\]

\[
= \inf_{\lambda \in C} \frac{1}{2\gamma \lambda} \left( 2\langle z^k - z_\lambda, z_\lambda - x \rangle_U + \|z_\lambda - z^k\|^2_U + \left( \frac{\rho - \varepsilon}{\beta \lambda} \right) \|z_\lambda - z^k\|^2 \right)
\]

\[
\leq \inf_{\lambda \in C} \frac{1}{2\gamma \lambda} \left( \frac{\rho - \varepsilon}{\beta \lambda} \right) \|z_\lambda - z^k\|^2_U \leq \frac{1}{\gamma \lambda} \langle z^k - z_\lambda, z_\lambda - x \rangle_U \leq \frac{\|z^0 - z^*\|_U + \|z^* - x\|_U}{\gamma \sqrt{\tau(k + 1)}} \frac{\|z^0 - z^*\|_U}{\gamma \sqrt{\tau(k + 1)}}. \tag{4.4.4}
\]

165
Part 3: We prove the result in the case that \( f \) is Lipschitz because the other case is symmetric. Proposition 4.2.4 proves the following identity: 
\[
\kappa^k_\alpha(\lambda) = (1 - 2/\lambda) \|z_\lambda - z^k\|_U^2.
\]
Thus, the fundamental inequality, the cosine rule, and the identities \( x^k_f - x^k_g = (1/\lambda)(z_\lambda - z^k) = T z^k - z^k, \)
\( x^k_g - x^k_S = w(x^k_g - x^k_f), \)
and \( C = (0, 2] \) show
\[
G^{\text{pre}}(x^k_g, x^k_g, x^k_S; x) 
\leq G^{\text{pre}}(x^k_f, x^k_g, x^k_S; x) + f(x^k_g) - f(x^k_f) + (S(x^k_g - x^k_S), -x)
\leq \inf_{\lambda \in C} \frac{1}{2\gamma \lambda} \left( \left(1 - \frac{2}{\lambda}\right) \|z_\lambda - z^k\|_U^2 + \|z^k - x\|^2_U - \|z_\lambda - x\|^2_U \right)
+ L \|x^k_g - x^k_f\| + \|w\|\|S\|\|x^k_g - x^k_f\|\|x\|
\overset{(4.1.7)}{=} \inf_{\lambda \in C} \frac{1}{2\gamma \lambda} \left( 2\langle z^k - z_\lambda, z_\lambda - x \rangle_U + 2 \left(1 - \frac{1}{\lambda}\right) \|z_\lambda - z^k\|_U^2 \right)
+ L \|x^k_g - x^k_f\| + \|w\|\|S\|\|x^k_g - x^k_f\|\|x\|
\overset{(4.4.4)}{\leq} \frac{(\|z^0 - z^*\|_U + \|z^* - x\|_U)\|z^0 - z^*\|_U}{\gamma \sqrt{Z(k + 1)}}
+ \frac{(L + \|w\|\|S\|\|x\|)\|z^0 - z^*\|_U}{\rho \sqrt{Z(k + 1)}}.
\]
\[
\]
\[
\]

\[
\]

Remark 4.4.1. Note that we can immediately strengthen the convergence result for PRS in Theorems 4.4.2 and 4.3.1. Indeed, we only need to assume that \( f \) or \( g \) is Lipschitz on the closed ball \( \overline{B}_U(x^*; \|z^0 - z^*\|_U \) of radius \( \|z^0 - z^*\|_U \) (under the metric \( \|\cdot\|_U \)) because for all \( k \geq 0, \)
\[
\|x^k_g - x^*\|_U^2 = \|J_{U^{-1}(\partial g + (1 - w)S)}(z^k) - J_{U^{-1}(\partial g + (1 - w)S)}(z^*)\|_U \leq \|z^k - z^*\|_U
\leq \|z^0 - z^*\|_U
\]
and, by a similar derivation, \( \|x^k_f - x^*\|_U \leq \|z^0 - z^*\|_U. \) Thus, the sequences lie in the ball:
\( (x^j_f)_{j \geq 0}, (x^j_g)_{j \geq 0} \subseteq B(x^*, \|z^0 - z^*\|_U \). \) We also have \( (x^j_g)_{j \geq 0}, (x^j_g)_{j \geq 0} \subseteq B(x^*, \|z^0 - z^*\|_U \) by the convexity of the ball. See [11, Proposition 8.28] for conditions that ensure Lipschitz continuity of convex functions on balls.
Remark 4.4.2. Note that the PRS rates in Theorem 4.4.2 above are sharp (up to constants). See Theorem 2.7.2 of Chapter 2 for the result when $S$ is the zero map.

Remark 4.4.3. In general, it is infeasible to take the supremum over the last component of $G_{pre}$ as in Equation (4.2.9). Thus, in practice we cannot use the pre-primal-dual gap to measure convergence. However, Theorem 4.4.2 bounds the pre-primal-dual gap at the $k$-th iteration by a multiple of the expression $\|Tz^k - z^k\| \|x\|$. Thus, if the supremum in Equation (4.2.9) can be restricted to a bounded set $D$, then $\|Tz^k - z^k\| \sup_{x \in D} \|x\|$ can be used as a proxy for the size of the pre-primal-dual gap. See section 4.5.1 for examples of such sets $D$.

4.5 Applications

In this section we will show that the four algorithms from Section 4.2.2 are capable of solving highly structured optimization problems:

Problem 4.5.1 (Model problem). Let $\mathcal{H}_0$ be a Hilbert space, and let $f, g : \Gamma_0(\mathcal{H}_0)$. For $i = 1, \cdots, n$, let $\mathcal{H}_i$ be a Hilbert space, let $h_i, l_i \in \Gamma_0(\mathcal{H}_i)$, suppose that $h_i \square l_i \in \Gamma_0(\mathcal{H}_i)$, and let $B_i : \mathcal{H}_0 \to \mathcal{H}_i$ be a bounded linear map. Finally, let $B : \mathcal{H}_0 \to \prod_{i=1}^n \mathcal{H}_i$ be the map $x \mapsto (B_1x, \cdots, B_nx)$. Then our model problem is as follows:

$$\min_{x \in \mathcal{H}_0} f(x) + g(x) + \sum_{i=1}^n (h_i \square l_i)(B_i x).$$

(4.5.1)

In addition, the dual problem is to

$$\min_{y \in \prod_{i=1}^n \mathcal{H}_i} (f^* \square g^*)(-B^*y) + \sum_{i=1}^n (h_i + l_i)(y_i).$$

All of the algorithms we consider take full advantage of the structure of the infimal convolution in Problem 4.5.1. We note that infimal convolutions are not wide spread in applications. Generally, we think of $h_i \square l_i$ as a regularization of $h_i$ by $l_i$, or vice versa. Indeed, under mild conditions, the smoothness of at least one of $h_i$ and $l_i$ implies the smoothness of the infimal convolution [11, Section 18.3]. When $l_i$ or $h_i$ is chosen properly,
this operation is sometimes called dual-smoothing [95]. Finally, we note that we can remove the infimal convolution operation from Problem 4.5.1 by setting $l_i = \iota_{\{0\}}$ because $h_i \square l_i = h_i$ for all $i = 1, \cdots, n$.

We assume the existence of a specific type of solution of Problem 4.5.1.

**Assumption 4.5.1.** *We assume that there exists*

$$x^* \in \text{zer} \left( \partial f + \partial g + \sum_{i=1}^{n} B_i^* (\partial h_i \square \partial l_i)(B_i(\cdot)) \right).$$

See [11, Proposition 16.5] or [47, Proposition 4.3] for conditions that guarantee the existence of $x^*$. In general, the containment

$$\text{zer} \left( \partial f + \partial g + \sum_{i=1}^{n} B_i^* (\partial h_i \square \partial l_i)(B_i(\cdot)) \right) \subseteq \text{zer} \left( \partial \left( f + g + \sum_{i=1}^{n} (h_i \square l_i)(B_i(\cdot)) \right) \right)$$

always holds, but the sets may not be equal. Nevertheless, this assumption is standard.

We now review two possible splittings of Problem 4.5.1. Both splittings will be designated by a “level.” The level is an indication of the number of extra dual variables that are introduced into the problem. Introducing more dual variables makes the problem further separable, and, hence, further parallelizable, but it also increases the memory footprint of the algorithm. It is unclear whether the number of dual variables affects the practical convergence speed of the algorithm in a negative way.

The following proposition is a simple exercise in duality, so we omit the proof.

**Proposition 4.5.1** (Level 1 optimality conditions). *Let $H = \prod_{i=0}^{n} \mathcal{H}_i$, and denote an arbitrary point $x \in H$ by $x = (x_1, y_1, \cdots, y_n) = (x, y)$. Let $f(x) = f(x) + \sum_{i=1}^{n} h_i^*(y_i)$, let $g(x) = g(x) + \sum_{i=1}^{n} l_i^*(y_i)$, and let $S : H \to H$ be the skew map $(x, y) \mapsto (B^*y, -Bx)$. Then

$$0 \in \partial f(x^*) + \partial g(x^*) + \sum_{i=1}^{n} B_i^* (\partial h_i \square \partial l_i)(B_i x^*)$$

(4.5.2) if, and only if, there is a vector $y^* \in \prod_{i=1}^{n} \mathcal{H}_i$ such that

$$0 \in \partial f(x^*, y^*) + \partial g(x^*, y^*) + S(x^*, y^*).$$

(4.5.3)
Notice that the subdifferential operators $\partial f$ and $\partial g$ in Equation (4.5.3) are completely separable in the variables of the product space $\mathbf{H}$. Thus, evaluating the proximity operators of $f$ and $g$ can be quite simple. However, the resolvent $J_{\partial f+\partial g}$ is not necessarily simple to evaluate. This difficulty motivates the introduction of new metrics on $\mathbf{H}$ that simplify the resolvent computation (Section 4.5.2).

Whenever the functions $l_i^*$ are Lipschitz differentiable, or equivalently, $l_i$ is strongly convex [11, Theorem 18.15], we can apply FBS or FBF (Algorithms 8 and 10) to the splitting in Proposition 4.5.1. For nonsmooth $l_i^*$, we can apply the PRS algorithm.

The proof of the following proposition is similar to Proposition 4.5.1, so we omit it. The proposition is most useful in the case that $l_i^*$ is not differentiable.

**Proposition 4.5.2** (Level 2 optimality conditions). Let $\mathbf{H} = \mathcal{H}_0 \times (\prod_{i=1}^n \mathcal{H}_i)^2$, and denote an arbitrary $x \in \mathbf{H}$ by $x = (x, y_1, \cdots, y_n, v_1, \cdots, v_n) = (x, y, v)$. Let $f(x) = f(x) + \sum_{i=1}^n (h_i^*(y_i) + l_i(v_i))$, let $g(x) = g(x)$, let $B : \mathcal{H}_0 \to \prod_{i=0}^n \mathcal{H}_i$ be the map $x \mapsto (B_1 x, \cdots, B_n x)$, and let $S : \mathbf{H} \to \mathbf{H}$ be the skew map $(x, y, v) \mapsto (B^* y, -B x + v, -y)$.

Then

$$0 \in \partial f(x^*) + \partial g(x^*) + \sum_{i=1}^n B_i^* (\partial h_i \square \partial l_i)(B_i x^*), \quad (4.5.4)$$

if, and only if, there is a vector $(y^*, v^*) \in (\prod_{i=1}^n \mathcal{H}_i)^2$ such that

$$0 \in \partial f(x^*, y^*, v^*) + \partial g(x^*, y^*, v^*) + S(x^*, y^*, v^*). \quad (4.5.5)$$

Note that if $l_i$ is differentiable, we can “assign” it to the function $g$, instead of “assigning” it to $f$. If $g$ is also differentiable, we can apply FBS to the inclusion.

There are many splittings that solve Problem 4.5.1. Furthermore, the complexity of Problem 4.5.1 can be increased in various ways, e.g., by precomposing each of $h_i$ and $l_i$ with linear operators [15, 24], or by solving systems of such inclusions [44, 22]. We choose to discuss this relatively simple formulation for clarity of exposition.

The next several sections relate the results and notation of the previous sections to the level 1 and 2 splittings.
4.5.1 Primal-dual gap functions

In this section, we discuss the pre-primal-dual gap function in the context of the level 1 splitting in Proposition 4.5.1. We give sufficient conditions for the gap function (Definition 4.2.1) to bound the primal and dual objectives of Problem 4.5.1 and show that the pre-primal-dual gap also bounds certain squared norms that arise from the strong convexity and differentiability of the terms of the objective.

In the level 1 splitting, the pre-primal-dual gap has the following form: for all \((x, y), (x^*, y^*) \in H\), we have

\[
G_{\text{pre}}(x, x; x^*) = f(x) + g(x) - f(x^*) - g(x^*) + \langle x - x^*, B^* y^* \rangle
\]

\[
+ \sum_{i=1}^{n} \left( h_i^*(y_i) + l_i^*(y_i) - h_i^*(y_i^*) - l_i^*(y_i^*) \right) - \langle B x^*, y - y^* \rangle,
\] (4.5.6)

where we used the identity \(\langle Sx, -x^* \rangle = \langle Sx, x - x^* \rangle\). If \(x^*\) is optimal for the inclusion in Proposition 4.5.1, then \(B^* y^* \in \partial f(x^*) + \partial g(x^*)\) and \(B_i x^* \in \partial h_i^*(y_i^*) + \partial l_i^*(y_i^*)\). (4.5.7)

We will now bound several terms that arise from the strong convexity and Lipschitz differentiability of the terms in the objective function.

We follow the convention that every closed, proper, and convex function \(F : H_0 \to (-\infty, \infty]\) is \(\mu_F\)-strongly convex and \(\tilde{\nabla} F\) is \((1/\beta_F)\)-Lipschitz for nonnegative real numbers \(\mu_F \geq 0\) and \(\beta_F \geq 0\). (If \(\beta_F = 0\), then \(F\) is not differentiable.) Note that we allow the constants \(\beta_F\) and \(\mu_F\) to vanish. If \(\beta_F > 0\), then \(\tilde{\nabla} F = \nabla F\) is Lipschitz. The following quantity is useful for summarizing the lower bounds that we derive from strong convexity and Lipschitz differentiability: for all \(x, y \in \text{dom}(F)\), if

\[
S_F(x, y) := \max \left\{ \frac{\mu_F}{2} \|x - y\|^2, \frac{\beta_F}{2} \|\tilde{\nabla} F(x) - \tilde{\nabla} F(y)\|^2 \right\},
\] (4.5.8)

then [11, Theorem 18.15]

\[
F(x) \geq F(y) + \langle x - y, \tilde{\nabla} F(y) \rangle + S_F(x, y).
\] (4.5.9)
We use the analogous notation for \( f, g \) and the conjugate functions \( h_i^*, l_i^* \) for \( i = 1, \ldots, n \).

Therefore, if we apply the lower bound in Equation (4.5.9) to each of the functions in Equation (4.5.6) and use the subgradient identities in Equation (4.5.7) to cancel inner products, we get

\[
G^{\text{pre}}(x, x, x; x^*) \geq S_f(x, x^*) + S_g(x, x^*) + \sum_{i=1}^{n} \left( S_{h_i^*}(y_i, y_i^*) + S_{l_i^*}(y_i, y_i^*) \right).
\] (4.5.10)

Equation (4.5.10) shows that convergence rates for the pre-primal-dual gap function immediately imply the same convergence rates for the \( S(\cdot, \cdot) \) functions in Equation (4.5.8).

Note that this lower bound does not require that \( \text{dom}(f) \) or \( \text{dom}(g) \) are bounded.

The next proposition gives sufficient conditions under which the pre-primal-dual gap bounds the primal and dual objectives. In general, we cannot expect such a bound to hold, unless several terms in the objective are Lipschitz continuous or certain subdifferentials are locally bounded.

**Proposition 4.5.3** (Level 1 gap function bounds). Let \( x^* \) be a minimizer of Problem 4.5.1. Assume the notation of Proposition 4.5.1. Let \( D_1 \subseteq \mathcal{H} \) and let \( D_2 \subseteq \prod_{i=1}^{n} \mathcal{H}_i \) be bounded sets. Then for any sequence of points \( ((x^j, y^j))_{j \geq 0} \subseteq \text{dom}(f + g) \times \prod_{i=1}^{n} \text{dom}(h_i^* + l_i^*) \), the inequality

\[
\begin{align*}
f(x^k) + g(x^k) + \sum_{i=1}^{n} (h_i \Box l_i)(B_ix^k) &- \left( f(x^*) + g(x^*) + \sum_{i=1}^{n} (h_i \Box l_i)(B_ix^*) \right) \\
\leq \sup_{x \in \{x^*\} \times D_2} G^{\text{pre}}(x^k, x^k, x^k, x)
\end{align*}
\]

holds for all \( k \geq 0 \) provided either of the following hold:

1. \( \text{dom}(h_1^* + l_1^*) \times \cdots \times \text{dom}(h_n^* + l_n^*) \subseteq D_2 \);
2. \( \partial(h_1 \Box l_1)(B_1x^k) \times \cdots \times \partial(h_n \Box l_n)(B_nx^k) \subseteq D_2 \).

Similarly, the inequality

\[
\begin{align*}
(f^* \Box g^*)(-B^*y^k) + \sum_{i=1}^{n} (h_i^* + l_i^*)(y_i^k) &- \left( (f^* \Box g^*)(-B^*y^*) + \sum_{i=1}^{n} (h_i^* + l_i^*)(B_iy_i^*) \right) \\
\leq \sup_{x \in D_1 \times \{y^*\}} G^{\text{pre}}(x^k, x^k, x^k, x)
\end{align*}
\]

171
holds for all \( k \geq 0 \) provided either of the following hold:

1. \( \text{dom}(f + g) \subseteq D_1 \);

2. \( \partial (f^* \Box g^*)(-B^*y^k) \subseteq D_1 \).

**Proof.** We only consider the primal case because the dual case is similar. For all \( i \in \{1, \ldots, n\} \) and \( k \geq 0 \), the Fenchel-Moreau Theorem [11, Theorem 13.32], the identity \( h_i \Box l_i = (h_i^* + l_i^*)^* \), and Conditions 1 and 2 show that we can reduce the domain of the following supremum:

\[
\sum_{i=1}^{n} (h_i \Box l_i)(B_i x^k) = \sup_{y \in \mathbb{H}} \left( \langle Bx^k, y \rangle - \sum_{i=1}^{n} (h_i^*(y_i) + l_i^*(y_i)) \right) \\
= \sup_{y \in D_2} \left( \langle Bx^k, y \rangle - \sum_{i=1}^{n} (h_i^*(y_i) + l_i^*(y_i)) \right).
\]

In addition, the Fenchel-Young inequality shows that

\[
\sum_{i=1}^{n} (h_i^*(y_i^k) + l_i^*(y_i^k)) - \langle x^*, B^*y^k \rangle \geq - \sum_{i=1}^{n} (h_i \Box l_i)(B_i x^*).
\]

Therefore,

\[
\sup_{x \in \{x^*\} \times D_2} G_{\text{pre}}(x^k, x^k, x^k; x) \\
= f(x^k) + g(x^k) - f(x^*) - g(x^*) + \sum_{i=1}^{n} h_i^*(y_i^k) + l_i^*(y_i^k) - \langle x^*, B^*y^k \rangle \\
+ \sup_{y \in D_2} \left( \langle Bx^k, y \rangle - \sum_{i=1}^{n} (h_i^*(y_i) + l_i^*(y_i)) \right) \\
\geq f(x^k) + g(x^k) + \sum_{i=1}^{n} (h_i \Box l_i)(B_i x^k) - \left( f(x^*) + g(x^*) + \sum_{i=1}^{n} (h_i \Box l_i)(B_i x^*) \right).
\]

\[\square\]

The bounded domain conditions in Proposition 4.5.3 are related to the Lipschitz continuity of the objective functions. Indeed, if \( h_i \) is Lipschitz, it follows that \( \text{dom}(h_i^*) \) is bounded [19, Proposition 4.4.6]. In addition, \( \text{dom}(h_i^* + l_i^*) = \text{dom}(h_i^*) \cap \text{dom}(l_i^*) \). Thus, if \( h_i^* \) has bounded domain, so does \( h_i^* + l_i^* \).
The bounded subgradient conditions in Proposition 4.5.3 are satisfied for \( h_i \square l_i \) if the infimal convolution is continuous everywhere and the sequence \( (B_i x^j)_{j \geq 0} \) is convergent. Indeed, in this case \( \partial(h_i \square l_i) \) is locally bounded \cite[Proposition 16.14(iii)]{11} and, hence, the union \( \bigcup_{j \geq 0} \partial(h_i \square l_i)(B_i x^j) \) is bounded. See \cite[Remark 2.2]{25} for similar remarks for a primal-dual forward-backward-forward splitting algorithm.

### 4.5.2 Two algorithm classes

In this section, we study the algorithms that arise for different classes of maps \((U_j)_{j \geq 0}\) and show how to compute the resolvent and forward-backward operators needed in order to apply the PPA, FBS, PRS, and FBF algorithms just as they appear in Section 4.2.

We fix the following notation for the rest of this section: Let \( \mu_{V_i} > 0 \) and let \( V_i \in S_{\mu_{V_i}}(H_i) \) for \( i = 0, \ldots, n \). Let \( \mu_{W_i} > 0 \) and let \( W_i \in S_{\mu_{W_i}}(H_i) \) for \( i = 1, \ldots, n \). These strongly monotone maps induce metrics on the spaces \( H_i \) for \( i = 0, \ldots, n \). They can be as simple as “diagonal” metrics, but they can also incorporate second order information. A discussion on the best metric choice is beyond the scope of this chapter, so we just refer the reader to \cite{103} for some applications of fixed “diagonal” metrics, and \cite{67} for varying “diagonal” metrics that satisfy conditions akin to Assumption 4.1.3.

Now define “block-diagonal” map

\[
V := V_1 \oplus \cdots \oplus V_n \in S_{\mu_V} \left( \prod_{i=1}^n H_i \right) \quad \text{and} \quad W := W_1 \oplus \cdots \oplus W_n \in S_{\mu_W} \left( \prod_{i=1}^n H_i \right)
\]

\[(4.5.11)\]

where \( \mu_V = \min\{\mu_{V_1}, \ldots, \mu_{V_n}\} \), and \( \mu_W = \min\{\mu_{W_1}, \ldots, \mu_{W_n}\} \). The rest of this section will build three types of metrics from \( V_0, V, W \).

Finally, note that Part 1 of Proposition 4.1.1 shows the following: for all \( z \in H \),

\[
z^+ = J_{U^{-1}(\partial f + S)}(z) \quad \iff \quad U(z - z^+) \in \partial f(z^+) + Sz^+.
\]

\[(4.5.12)\]

See Proposition 4.5.5, 4.5.7, and 4.5.8 for examples of resolvent computations.
4.5.2.1 First metric class

In this section, our metrics depend on a parameter $w$, which appears in Algorithm 9. We only use the metric for the case that $w \in \{0, 1/2, 1\}$, but we state all of our results for the general case $w \in \mathbb{R}$. The case $w = 1/2$ first appeared in [23, Theorem 2.1], and the case $w = 1$ first appeared in [71].

**Proposition 4.5.4.** Let $w \in \mathbb{R}$. Assume the setting of Proposition 4.5.1. For all $x = (x, y) \in H$, define the map:

$$U_w x := (V_0 x - w B^* y, -w B x + V y).$$

(4.5.13)

Suppose that $w^2 \| V^{-1/2} B V_0^{-1/2} \|^2 < 1$. Then $U_w$ is self adjoint and strongly monotone: for all $x \in H$,

$$\langle x, U_w x \rangle \geq \frac{1}{2} \left( 1 - w^2 \| V^{-1/2} B V_0^{-1/2} \|^2 \right) \min\{\mu_{V_0}, \mu_{V}\} \left( \|x\|^2 + \|y\|^2 \right).$$

(4.5.14)

Assume the setting of Proposition 4.5.2. For all $x = (x, y, v) \in H$, define the map:

$$U'_w x := (V_0 x - w B^* y, V y - w B x + w v, w y + W v).$$

(4.5.15)

Suppose that $w^2 \| V^{-1/2} B V_0^{-1/2} \|^2 + w^2 \| W^{-1/2} V^{-1/2} \|^2 < 1$. Then

$$\langle x, U'_w x \rangle \geq \frac{1}{3} \left( 1 - w^2 \| V^{-1/2} B V_0^{-1/2} \|^2 - w^2 \| W^{-1/2} V^{-1/2} \|^2 \right) \times \min\{\mu_{V_0}, \mu_{V}, \mu_{W}\} \left( \|x\|^2 + \|y\|^2 + \|v\|^2 \right).$$

(4.5.16)

We omit the proof of Proposition 4.5.4 because Equation (4.5.14) is shown in [101, Lemma 4.3, Equation (4.14)] when $w = 1$ and the extension to general $w$ is straightforward, and because Equation (4.5.16) has nearly the same proof.

Note that our conditions for ergodic convergence in Theorem 4.3.1 require the metric inducing maps to be almost decreasing up to a summable residual in the Loewner partial ordering $\succ$ (see Section 4.1.2). If $(U_j)_{j \geq 0}$ is a sequence of maps defined as in Equation (4.5.13), we have

$$(U_k - U_{k+1}) x = ((V_{0,k} - V_{0,k+1}) x, (V_k - V_{k+1}) y)$$
for all $x \in H$. Thus, if for all $k \geq 0$, we have $V_{0,k} \geq V_{0,k+1}$ and $V_k \geq V_{k+1}$, we can guarantee that the product metric is decreasing (Lemma 4.1.1). A similar result holds for the level 2 metrics in Equation (4.5.15).

The following proposition shows how to evaluate the FBS operator under the metrics induced by $U_w$ and $U'_w$. Note that the results of Proposition 4.5.5 are not new. The level 1 case with $w \in \{0, 1/2, 1\}$ has appeared implicitly in several papers, including [52, 112, 49]. It has also explicitly appeared in [101, Lemma 4.5]. In addition, the proof of the level 2 case appeared in [23, Equation (2.38)]. Thus, we omit the proof.

**Proposition 4.5.5** (Forward-Backward operators under the first metric class). Let $w \in \mathbb{R}$. Assume the setting of Proposition 4.5.1, and suppose that $U_w \in S_\rho(H)$ (Equation (4.5.13)) for some $\rho > 0$. Let $z := (x, y) \in H$. Suppose that $g,l^*_1, \cdots, l^*_n$ are differentiable. Then $z^+ := J_{U^{-1}_w(\partial f + wS)}(z - U^{-1}_w \nabla g(z))$ has the following form:

$$x^+ = \text{prox}^V_0(x - V^{-1}_0(wB^*y + \nabla g(x)));$$

for $l = 1, 2, \ldots, n$, in parallel do

$$y^+_i = \text{prox}^{V^*_i}_{h^*_i}(y_i + V^{-1}_i(wB_i(2x^+ - x) - \nabla l^*_i(y_i));$$

Assume the setting of Proposition 4.5.2, and suppose that $U'_w \in S_\rho(H)$ (Equation (4.5.15)) for some $\rho > 0$. Let $z := (x, y, v) \in H$, and suppose that $g$ is differentiable. Then $z^+ := J_{(U^{-1}_w)'(\partial f + wS)}(z - (U^{-1}_w)' \nabla g(z))$ has the following form:

$$x^+ = \text{prox}^V_j(x - V^{-1}_0(wB^*y + \nabla g(x));$$

for $l = 1, 2, \ldots, n$, in parallel do

$$v^+_i = \text{prox}^{W^*_i}_{l^*_i}(v_i + wW^{-1}_i y_i);$$

$$y^+_i = \text{prox}^{V^*_i}_{h^*_i}(y_i + V^{-1}_i(wB_i(2x^+ - x) - (2v^+_i - v_i));$$

### 4.5.3 Second metric class

The following proposition appeared implicitly in [101, Section 4.2, Equation (4.16)] for $w = 1$. We reproduce the result for completeness.
Proposition 4.5.6. Assume the setting of Proposition 4.5.1. For all $x \in H$, define:

$$U_w x := (V_0 x, (V - w^2 B V_0^{-1} B^*) y). \quad (4.5.17)$$

Suppose that $w^2 \|V^{-1/2} B V_0^{-1/2}\|^2 < 1$. Then $U$ is self adjoint and strongly monotone: for all $x \in H$,

$$\langle x, U_w x \rangle \geq \min \left\{ \mu V_0, \left( 1 - w^2 \|V^{-1/2} B V_0^{-1/2}\|^2 \right) \mu V \right\} \left( \|x\|^2 + \|y\|^2 \right). \quad (4.5.18)$$

Proof. Set $C = w B$. For all $y \in \prod_{i=1}^n H_i$, we have

$$\langle y, (V - CV_0^{-1} C^*) y \rangle = \langle V^{1/2} y, (I_{\prod_{i=1}^n H_i} - V^{-1/2} C V_0^{-1} C^* V^{-1/2}) V^{1/2} y \rangle$$

$$= \langle Vy, y \rangle - \langle V^{1/2} y, V^{-1/2} C V_0^{-1} C^* V^{-1/2} V^{1/2} y \rangle$$

$$\geq \left( 1 - \|V^{-1/2} C^* V_0^{-1} C V^{-1/2}\| \right) \langle Vy, y \rangle$$

$$\geq \left( 1 - w^2 \|V^{-1/2} B V_0^{-1/2}\|^2 \right) \mu V \|y\|.$$
The following proposition shows how to evaluate the FBS operator under the metric induced by $U$. Note that Proposition 4.5.7 appears in [101, Lemma 4.10] for $w = 1$. Thus, we omit the proof.

**Proposition 4.5.7** (Forward-Backward operators under the second metric class). Assume the setting of Proposition 4.5.1. Suppose that $f \equiv 0$, and that $U \in \mathcal{S}_\rho(H)$ (Equation (4.5.17)) for some $\rho > 0$. Let $z := (x, y) \in H$. Suppose that $g, l^*_1, \ldots, l^*_n$ are differentiable. Then $z^+ := J_{U^{-1}(\partial f + S)}(z - U^{-1}\nabla g(z))$ has the following form:

$$
\begin{align*}
&\text{for } l = 1, 2, \ldots, n, \text{ in parallel do} \\
&\quad y^+_i = \text{prox}_{h^*_i} (y_i + V^{-1}_i \left( wB_i \left( x - V_0^{-1}(\nabla g(x) + wB^*y) - \nabla l^*_i(y_i) \right) \right)) \\
&\quad x^+ = x - V_0^{-1}(\nabla g(x) + wB^*y^+) \\
\end{align*}
$$

Now consider the special case $w = 0$. In this case, the first and second metric classes agree. The following Proposition with $U = I_H$ appears in [29, Proposition 2.7]. Our generalization is straightforward, so we omit the proof.

**Proposition 4.5.8** (Resolvents of skew operators). Assume the setting of Proposition 4.5.1. Suppose that $U_w \in \mathcal{S}_\rho(H)$ (Equation (4.5.17)) for some $\rho > 0$. Let $z := (x, y) \in H$. Then $z^+ := J_{U^{-1}S}(z)$ has the following form:

$$
(x^+, y^+) = ((I_{H_0} + \gamma^2 V_0 B^*V)\gamma V_0 B^*)^{-1}(x - \gamma V_0 B^*)y, (I_{P_{\Pi_{i=1}^n H_i} + \gamma^2 VBV_0 B^*})^{-1}(y + \gamma VBx)).
$$

Generalizing the resolvent operator computation in Proposition 4.5.8 to the level 2 case is straightforward, though slightly messy. It has not found application in the literature yet, so we omit the statement.

### 4.5.4 New and old convergence rates

Table 4.5.4 lists the application of PPA, FBS, PRS, and FBF algorithms under the metrics introduced in Section 4.5.2 and indicates which convergence rates have been shown in the literature. We note that, to the best of our knowledge, for all of the methods we
<table>
<thead>
<tr>
<th>Reference</th>
<th>Algorithm</th>
<th>Metric</th>
<th>Level</th>
<th>$w$</th>
<th>Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>[36, Algorithm 1]</td>
<td>PPA</td>
<td>(4.5.13)</td>
<td>1</td>
<td>1</td>
<td>$O(1/(k + 1))$ ergodic [36]</td>
</tr>
<tr>
<td>[23, Algorithm 2.2]</td>
<td>PPA</td>
<td>(4.5.15)</td>
<td>2</td>
<td>1</td>
<td>none</td>
</tr>
<tr>
<td>[52, 112]</td>
<td>FBS</td>
<td>(4.5.13)</td>
<td>1</td>
<td>1</td>
<td>$O(1/(k + 1))$ ergodic [20]</td>
</tr>
<tr>
<td>[40, 45, 101]</td>
<td>FBS</td>
<td>(4.5.17)</td>
<td>1</td>
<td>1</td>
<td>none</td>
</tr>
<tr>
<td>[23, Algorithm 2.1]</td>
<td>PRS</td>
<td>(4.5.13)</td>
<td>1</td>
<td>$1/2$</td>
<td>none</td>
</tr>
<tr>
<td>[29, Remark 2.9], [96]</td>
<td>PRS</td>
<td>(4.5.17)</td>
<td>1</td>
<td>0</td>
<td>none</td>
</tr>
<tr>
<td>[29, 47]</td>
<td>FBF</td>
<td>(4.5.17)</td>
<td>1</td>
<td>0</td>
<td>$O(1/(k + 1))$ ergodic [25]</td>
</tr>
</tbody>
</table>

**Table 4.1:** This table lists the original appearance of the algorithms constructed from pairing the metrics in Section 4.5.2 with the PPA, FBS, PRS, and FBF algorithms applied to Problem 4.5.1. See Propositions 4.5.1 and 4.5.2 for the definitions of the “level.”

discuss, the nonergodic fixed metric convergence rates, the ergodic convergence rates under variable metrics, and the nonergodic/ergodic convergence rates with nonconstant relaxation have never appeared in the literature.

Any pairing between metrics, algorithms, and splittings that does not appear in Table 4.5.4 is an algorithm where, to the best of our knowledge, no convergence rate has appeared in the literature.

### 4.6 Conclusion

In this chapter, we provided a convergence rate analysis of a general monotone inclusion problem under the application of four different algorithms. We provided *ergodic* convergence rates under variable metrics, stepsizes, and relaxation, and recovered several known
rates in the process. In addition, for three of the algorithms we provided the first nonergodic primal-dual gap convergence rates that have appeared in the literature. Finally, we showed how our results imply convergence rates of a large class of primal-dual splitting algorithms. The techniques developed in this chapter are not limited to the four algorithms we chose to study, and the proofs of this chapter can be used as a template for proving convergence rates of other special cases of the unifying scheme.
CHAPTER 5

Convergence Rate Analysis of the Forward-Douglas-Rachford Splitting Scheme

5.1 Introduction

Operator-splitting schemes are algorithms for splitting complicated problems arising in PDE, monotone inclusions, optimization, and control into many simpler subproblems. The achieved decomposition can give rise to inherently parallel and, in some cases, distributed algorithms. These characteristics are particularly desirable for large-scale problems that arise in machine learning, finance, control, image processing, and PDE [26].

In optimization, the Douglas-Rachford splitting (DRS) algorithm [85] minimizes sums of (possibly) nonsmooth functions $f, g : \mathcal{H} \rightarrow (-\infty, \infty]$ on a Hilbert space $\mathcal{H}$:

$$\minimize_{x \in \mathcal{H}} f(x) + g(x). \quad (5.1.1)$$

During each step of the algorithm, DRS applies the proximal operator, which is the basic subproblem in nonsmooth minimization, to $f$ and $g$ individually rather than to the sum $f + g$. Thus, the key assumption in DRS is that $f$ and $g$ are easy to minimize independently, but the sum $f + g$ is difficult to minimize. We note that many complex objectives arising in machine learning [26] and signal processing [46] are the sum of nonsmooth terms with simple or closed-form proximal operators.

The forward-backward splitting (FBS) algorithm [100] is another technique for solving (5.1.1) when $g$ is known to be smooth. In this case, the proximal operator of $g$ is never evaluated. Instead, FBS combines gradient (forward) steps with respect to $g$ and
proximal (backward) steps with respect to $f$. FBS is especially useful when the proximal operator of $g$ is complex and its gradient is simple to compute.

Recently, the forward-Douglas-Rachford splitting (FDRS) algorithm [30] was proposed to combine DRS and FBS and extend their applicability (see Algorithm 11). More specifically, let $V \subseteq \mathcal{H}$ be a closed vector space and suppose $g$ is smooth. Then FDRS applies to the following constrained problem:

$$
\minimize_{x \in V} f(x) + g(x).
$$

(5.1.2)

Throughout the course of the algorithm, the proximal operator of $f$, the gradient of $g$, and the projection operator onto $V$ are all employed separately.

The FDRS algorithm can also apply to affinely constrained problems. Indeed, if $V = V_0 + b$ for a closed vector subspace $V_0 \subseteq \mathcal{H}$ and a vector $b \in \mathcal{H}$, then Problem (5.1.2) can be reformulated as

$$
\minimize_{x \in V_0} f(x + b) + g(x + b).
$$

(5.1.3)

For simplicity, we only consider linearly constrained problems.

The FDRS algorithm is a generalization of the generalized forward-backward splitting (GFBS) algorithm [105], which solves the problem minimize$_{x \in \mathcal{H}} \sum_{i=1}^{n} f_i(x) + g(x)$ where $f_i : \mathcal{H} \to (-\infty, \infty]$ are closed, proper, convex and (possibly) nonsmooth. In the GFBS algorithm, the proximal mapping of each function $f_i$ is evaluated in parallel. We note that GFBS can be derived as an application of FDRS to the equivalent problem:

$$
\min_{(x_1, x_2, \ldots, x_n) \in \mathcal{H}^n} \sum_{i=1}^{n} f_i(x_i) + g \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right).
$$

(5.1.4)

In this case, the vector space $V = \{(x, \ldots, x) \in \mathcal{H}^n \mid x \in \mathcal{H}\}$ is the diagonal set of $\mathcal{H}^n$ and the function $f$ is separable in the components of $(x_1, \ldots, x_n)$.

The FDRS algorithm is the only primal operator-splitting method capable of using all structure in Equation (5.1.2). In order to achieve good practical performance, the other primal splitting methods require stringent assumptions on $f$, $g$, and $V$. Primal DRS
cannot use the smooth structure of \( g \), so the proximal operator of \( g \) must be simple. On the other hand, primal FBS and forward-backward-forward splitting (FBFS) [111] cannot separate the coupled nonsmooth structure of \( f \) and \( V \), so minimizing \( f(x) \) subject to \( x \in V \) must be simple. In contrast, FDRS achieves good practical performance if it is simple to minimize \( f \), evaluate \( \nabla g \), and project onto \( V \).

Modern primal-dual splitting methods [36, 61, 52, 112, 29, 78] can also decompose problem (5.1.2), but they introduce extra variables and are, thus, less memory efficient. It is unclear whether FDRS will perform better than primal-dual methods when memory is not a concern. However, it is easier to choose algorithm parameters for FDRS and, hence, it can be more convenient to use in practice.

**Application: constrained quadratic programming and support vector machines**

Let \( d \) and \( m \) be natural numbers. Suppose that \( Q \in \mathbb{R}^{d \times d} \) is a symmetric positive semi-definite matrix, \( c \in \mathbb{R}^d \) is a vector, \( C \subseteq \mathbb{R}^d \) is a constraint set, \( A \in \mathbb{R}^{m \times d} \) is a linear map, and \( b \in \mathbb{R}^m \) is a vector. Consider the problem:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \langle Qx, x \rangle + \langle c, x \rangle \\
\text{subject to:} & \quad x \in C \\
& \quad Ax = b.
\end{align*}
\]  

Problem (5.1.5) arises in the dual form soft-margin kernelized support vector machine classifier [54] in which \( C \) is a box constraint, \( b \) is 0, and \( A \) has rank one. Note that by the argument in (5.1.3), we can always assume that \( b = 0 \).

Define the smooth function \( g(x) = (1/2)\langle Qx, x \rangle + \langle c, x \rangle \), the indicator function \( f(x) = \chi_C(x) \) (which is 0 on \( C \) and \( \infty \) elsewhere), and the vector space \( V = \{ x \in \mathbb{R}^d \mid Ax = 0 \} \). With this notation, (5.1.5) is in the form (5.1.2) and, thus, FDRS can be applied. This splitting is nice because \( \nabla g(x) = Qx + c \) is simple whereas the proximal operator of \( g \) requires a matrix inversion \( \text{prox}_{\gamma g} = (I_{\mathbb{R}^d} + \gamma Q)^{-1} \circ (I_{\mathbb{R}^d} - \gamma c) \), which is expensive for
large-scale problems.

5.1.1 Goals, challenges, and approaches

This work seeks to characterize the convergence rate of the FDRS algorithm applied to Problem (5.1.2). Chapter 2 has shown that the sharp convergence rate of the fixed-point residual (FPR) (see Equation (5.1.23)) of the FDRS algorithm is $o\left(\frac{1}{k+1}\right)$. To the best of our knowledge, nothing is else is known about the convergence rate of FDRS. Furthermore, it is unclear how the FDRS algorithm relates to other algorithms. We seek to fill this gap.

The techniques used in this chapter are based on Chapters 2, 3 and, 4. These techniques are quite different from those used in classical objective error convergence rate analysis. The classical techniques do not apply because the FDRS algorithm is driven by the fixed-point iteration of a nonexpansive operator, not by the minimization of a model function. Thus, we must explicitly use the properties of nonexpansive operators in order to derive convergence rates for the objective error.

We summarize our contributions and techniques as follows:

i We analyze the objective error convergence rates (Theorems 5.3.1 and 5.3.2) of the FDRS algorithm under general convexity assumptions. We show that FDRS is, in the worst case, nearly as slow as the subgradient method yet nearly as fast as the proximal point algorithm (PPA) in the ergodic sense. Our nonergodic rates are shown by relating the objective error to the FPR through a fundamental inequality. We also show that the derived rates are sharp through counterexamples (Remarks 5.3.1 and 5.3.2).

ii We show that if $f$ or $g$ is strongly convex, then a natural sequence of points converges strongly to a minimizer. Furthermore, the best iterate converges with rate $o(1/(k + 1))$, the ergodic iterate converges with rate $O(1/(k + 1))$, and the nonergodic iterate converges with rate $o(1/\sqrt{k + 1})$. The results follow by showing that a certain sequence
of squared norms is summable. We also show that some of the derived rates are sharp by constructing a novel counterexample (Theorem 5.6.3).

iii We show that if \( f \) is differentiable and \( \nabla f \) is Lipschitz, then the best iterate of the FDRS algorithm has objective error of order \( o(1/(k+1)) \) (Theorem 5.5.1). This rate is an improvement over the sharp \( o(1/\sqrt{k+1}) \) convergence rate for nonsmooth \( f \). The result follows by showing that the objective error is summable.

iv We establish scenarios under which FDRS converges linearly (Theorem 5.6.1) and show that linear convergence is impossible under other scenarios (Theorem 5.6.3).

v We show that even if \( f \) and \( g \) are strongly convex, the FDRS algorithm can converge arbitrarily slowly (Theorem 5.6.2).

vi We show that the FDRS algorithm is the limiting case of a recently developed primal-dual forward-backward splitting algorithm (Section 5.7) and, thus, clarify how FDRS relates to existing algorithms.

Our analysis builds on the techniques and results of [30] and Chapters 2 and 3. The rest of this section contains a brief review of these results.

### 5.1.2 Notation and facts

Most of the definitions and notation that we use in this chapter are standard and can be found in [11]. Throughout this chapter, we use \( \mathcal{H} \) to denote (a possibly infinite dimensional) Hilbert space. In fixed-point iterations, \( (\lambda_j)_{j \geq 0} \subset \mathbb{R}_+ \) will denote a sequence of relaxation parameters, and

\[
\Lambda_k := \sum_{i=0}^k \lambda_i \quad (5.1.6)
\]

is its \( k \)th partial sum. Given the sequence \( (x^j)_{j \geq 0} \subset \mathcal{H} \), we let

\[
\bar{x}^k = \frac{1}{\Lambda_k} \sum_{i=0}^k \lambda_i x^i \quad (5.1.7)
\]
denote its kth average with respect to the sequence \((\lambda_j)_{j \geq 0}\). We call \((\pi_j)_{j \geq 0}\) the ergodic sequence and we call \((x_j)_{j \geq 0}\) the nonergodic sequence.

For any subset \(C \subseteq \mathcal{H}\), we define the distance function:

\[
d_C(x) := \inf_{y \in C} \|x - y\|. \tag{5.1.8}
\]

In addition, we define the indicator function \(\chi_C : \mathcal{H} \to \{0, \infty\}\) of \(C\): for all \(x \in C\) and \(y \in \mathcal{H} \setminus C\), we have \(\chi_C(x) = 0\) and \(\chi_C(y) = \infty\).

Given a closed, proper, and convex function \(f : \mathcal{H} \to (-\infty, \infty]\), the set \(\partial f(x) = \{p \in \mathcal{H} \mid \text{for all } y \in \mathcal{H}, f(y) \geq f(x) + \langle y - x, p \rangle\}\) denotes its subdifferential at \(x\) and \(\tilde{\nabla} f(x) \in \partial f(x)\) denotes a subgradient. (This notation was used in [16, Eq. (1.10)].) If \(f\) is Gâteaux differentiable at \(x \in \mathcal{H}\), we have \(\partial f(x) = \{\nabla f(x)\}\) [11, Proposition 17.26].

Let \(I_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}\) be the identity map on \(\mathcal{H}\). For any \(x \in \mathcal{H}\) and \(\gamma \in \mathbb{R}_{++}\), we let

\[
\text{prox}_{\gamma f}(x) := \arg \min_{y \in \mathcal{H}} \left( f(y) + \frac{1}{2\gamma} \|y - x\|^2 \right) \quad \text{and} \quad \text{refl}_{\gamma f} := 2\text{prox}_{\gamma f} - I_{\mathcal{H}},
\]

which are known as the proximal and reflection operators, respectively.

The subdifferential of the indicator function \(\chi_V\) where \(V \subseteq \mathcal{H}\) is a closed vector subspace is defined as follows: for all \(x \in \mathcal{H}\),

\[
\partial \chi_V(x) = \begin{cases} 
V^\perp & \text{if } x \in \mathcal{H}; \\
\emptyset & \text{otherwise} 
\end{cases} \tag{5.1.9}
\]

where \(V^\perp\) is the orthogonal complement of \(V\). Evidently, if \(P_V(\cdot) = \arg \min_{y \in V} \|y - \cdot\|^2\) is the projection onto \(V\), then

\[
\text{prox}_{\gamma \chi_V} = P_V \quad \text{and} \quad \text{refl}_{\gamma \chi_V} = 2P_V - I_{\mathcal{H}} = P_V - P_V^\perp,
\]

and these operators are independent of \(\gamma\).
Let \( \lambda > 0 \), let \( L \geq 0 \), and let \( T : \mathcal{H} \to \mathcal{H} \) be a map. The map \( T \) is called \( L \)-Lipschitz continuous if \( \|Tx - Ty\| \leq L\|x - y\| \) for all \( x, y \in \mathcal{H} \). The map \( T \) is called nonexpansive if it is 1-Lipschitz. We also use the notation:

\[
T_\lambda := (1 - \lambda)I_\mathcal{H} + \lambda T. \tag{5.1.10}
\]

If \( \lambda \in (0,1) \) and \( T \) is nonexpansive, then \( T_\lambda \) is called \( \lambda \)-averaged \cite[Definition 4.23]{11}.

We call the following identity the cosine rule:

\[
\|y - z\|^2 + 2\langle y - x, z - x \rangle = \|y - x\|^2 + \|z - x\|^2, \quad \forall x, y, z \in \mathcal{H}. \tag{5.1.11}
\]

Young’s inequality is the following: for all \( a, b \geq 0 \) and \( \varepsilon > 0 \), we have

\[
ab \leq a^2/(2\varepsilon) + \varepsilon b^2/2. \tag{5.1.12}
\]

5.1.3 Assumptions

**Assumption 5.1.1** (Convexity). \( f \) and \( g \) are closed, proper, and convex.

We also assume the existence of a particular solution to 5.1.2

**Assumption 5.1.2** (Solution existence). \( \text{zer}(\partial f + \nabla g + \partial \chi_V) \neq \emptyset \)

Finally we assume that \( \nabla g \) is sufficiently nice.

**Assumption 5.1.3** (Differentiability). The function \( g \) is differentiable, \( \nabla g \) is \((1/\beta)\)-Lipschitz, and \( P_V \circ \nabla g \circ P_V \) is \((1/\beta_V)\)-Lipschitz.

5.1.4 The FDRS algorithm

FDRS is summarized in Algorithm 11.

**Algorithm 11:** Relaxed Forward-Douglas-Rachford splitting (relaxed FDRS)

**input:** \( z^0 \in \mathcal{H}, \gamma \in (0, \infty), (\lambda_j)_{j\geq0} \in (0, \infty) \)

**for** \( k = 0, 1, \ldots \) **do**

\[
z^{k+1} = (1 - \lambda_k)z^k + \lambda_k \left( \frac{1}{2}I_\mathcal{H} + \frac{1}{2}\text{refl}_f \circ \text{refl}_\chi_V \right) \circ (I - \gamma P_V \circ \nabla g \circ P_V)(z^k);
\]
For now, we do not specify the stepsize parameters. See section 5.1.6 for choices that ensure convergence and, see Lemma 5.2.1 and Figure 5.1 for intuition.

Evidently, Algorithm 11 has the form: for all $k \geq 0$, $z^{k+1} = (T_{FDRS})_{\lambda_k}(z^k)$ where

$$T_{FDRS} := \left(\frac{1}{2}I_H + \frac{1}{2}\text{refl}_{\gamma f} \circ \text{refl}_{\chi V}\right) \circ (I - \gamma P_V \circ \nabla g \circ P_V).$$

Because $T_{FDRS}$ is nonexpansive (Part 7 of Theorem 5.1.1), it follows that the FDRS algorithm is a special case of the Krasnosel’ski˘ı-Mann (KM) iteration [79, 88, 43].

By choosing particular $f, g$ and $V$, we recover several other splitting algorithms:

DRS: ($g \equiv 0$) $z^{k+1} = (1 - \lambda_k)z^k + \lambda_k \left(\frac{1}{2}I_H + \frac{1}{2}\text{refl}_{\gamma f} \circ \text{refl}_{\chi V}\right)(z^k)$;

FBS: ($V = H$) $z^{k+1} = (1 - \lambda_k)z^k + \lambda_k \text{prox}_{\gamma f} \circ (I - \gamma \nabla g)(z^k)$;

FBS: ($f \equiv 0$) $z^{k+1} = (1 - \lambda_k)z^k + \lambda_k P_V \circ (z - \gamma P_V \circ \nabla g \circ P_V)(z^k)$.

For general $f, g$ and $V$, the primal DRS and FBS algorithms are not capable splitting Problem (5.1.2) in the same way as (5.1.13). Indeed, the DRS algorithm cannot use the smooth structure of $g$, and the FBS algorithm requires the evaluation of $\text{prox}_{\gamma (f + \chi V)}(\cdot) = \arg \min_{x \in V} (f(x) + (1/2\gamma)\|x - \cdot\|^2)$. The FDRS algorithm eliminates these difficult problems and replaces them with (possibly) more tractable ones.

5.1.5 Proximal, averaged, and FDRS operators

We briefly review some operator-theoretic properties.

**Proposition 5.1.1.** Let $\lambda > 0$, let $\gamma > 0$, let $\alpha > 0$, and let $f : H \to (-\infty, \infty]$ be closed, proper, and convex.

1. Optimality conditions of $\text{prox}$: Let $x \in H$. Then $x^+ = \text{prox}_{\gamma f}(x)$ if, and only if, $\tilde{\nabla}f(x^+) := (1/\gamma)(x - x^+) \in \partial f(x^+)$.  

2. Optimality conditions of $\text{prox}_{\chi_V}$: Let $x \in H$. Then $x^+ = \text{prox}_{\gamma \chi V}(x)$ if, and only if, $\tilde{\nabla}\chi_V(x^+) := (1/\gamma)(x - x^+) \in \partial \chi_V(x^+)$. Also, $\gamma \tilde{\nabla}\chi_V(x^+) = P_V x \in V^\perp$. 

187
3. Averaged operator contraction property: A map $T : \mathcal{H} \to \mathcal{H}$ is $\alpha$-averaged (see (5.1.10)) if, and only if, for all $x, y \in \mathcal{H}$,
\[
\|Tx - Ty\|^2 \leq \|x - y\|^2 - \frac{1 - \alpha}{\alpha} \|(I_{\mathcal{H}} - T)x - (I_{\mathcal{H}} - T)y\|^2.
\] (5.1.14)

4. Composition of averaged operators: Let $\alpha_1, \alpha_2 \in (0, 1)$. Suppose $T_1 : \mathcal{H} \to \mathcal{H}$ and $T_2 : \mathcal{H} \to \mathcal{H}$ are $\alpha_1$ and $\alpha_2$-averaged operators, respectively. Then for all $x, y \in \mathcal{H}$, the map $T_1 \circ T_2 : \mathcal{H} \to \mathcal{H}$ is averaged with parameter
\[
\alpha_{1,2} := \frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{1 - \alpha_1\alpha_2} \in (0, 1)
\] (5.1.15)

5. Wider relaxations: A map $T : \mathcal{H} \to \mathcal{H}$ is $\alpha$-averaged if, and only if, $T_{\lambda}$ (see (5.1.10)) is $\lambda\alpha$-averaged for all $\lambda \in (0, 1/\alpha)$.

6. Proximal operators are (1/2)-averaged: The operator $\text{prox}_{\gamma f} : \mathcal{H} \to \mathcal{H}$ is (1/2)-averaged and, hence, the operator $\text{refl}_{\gamma f} = 2\text{prox}_{\gamma f} - I_{\mathcal{H}}$ is nonexpansive.

7. Averaged property of the FDRS operator: Suppose that $\gamma \in (0, 2\beta)$. Then the operator $T_{\text{FDRS}}$ (see (5.1.13)) is $\alpha_{\text{FDRS}} := 2\beta/(4\beta - \gamma)$ averaged.

Proof. Parts 1, 2, 3, 5, and 6 can be found in [11]. Part 4 can be found in [50]. Part 7 follows from two facts: The operator $((1/2)I_{\mathcal{H}} + (1/2)\text{refl}_{\gamma f} \circ \text{refl}_{\chi V})$ is (1/2)-averaged by Part 6, and $I - \gamma P_V \circ \nabla g \circ P_V$ is $(\gamma/2\beta)$-averaged by [30, Proposition 4.1 (ii)]. Thus, Part 4 proves Part 7.

Remark 5.1.1. Later we require $(\lambda_j)_{j \geq 0} \subseteq (0, 1/\alpha_{\text{FDRS}})$ so we hope that $\alpha_{\text{FDRS}}$ is small. Note that the expression for $\alpha_{\text{FDRS}}$ is new and improves upon the previous constant: $\max\{2/3, 2\gamma/(\gamma + 2\beta)\}$. See also [50, Remark 2.7 (i)].

The proof of the following result is essentially contained in [50, Theorem 2.4]. We reproduce it in Appendix B.1 in order to derive a bound.

**Proposition 5.1.2.** Let $\alpha_1, \alpha_2 \in (0, 1)$. Suppose that $T_1 : \mathcal{H} \to \mathcal{H}$ and $T_2 : \mathcal{H} \to \mathcal{H}$ are $\alpha_1$ and $\alpha_2$-averaged operators, respectively, and that $z^*$ is a fixed-point of $T_1 \circ T_2.$
Define $\alpha_{1,2} \in (0, 1)$ as in (5.1.15). Let $z^0 \in \mathcal{H}$, let $\varepsilon \in (0, 1)$, and consider a sequence $(\lambda_j)_{j \geq 0} \subseteq (0, (1 - \varepsilon)(1 + \varepsilon \alpha_{1,2})/\alpha_{1,2})$. Let $(z^i)_{j \geq 0}$ be generated by the following iteration: for all $k \geq 0$, let $z^{k+1} = (T_1 \circ T_2)_{\lambda_k}(z^k)$. Then

$$\sum_{i=0}^{\infty} \lambda_i \|(I_{\mathcal{H}} - T_2)(z^i) - (I_{\mathcal{H}} - T_2)(z^*)\|^2 \leq \frac{\alpha_2(1 + 1/\varepsilon)\|z^0 - z^*\|^2}{1 - \alpha_2}.$$

**Remark 5.1.2.** Let $\varepsilon \in (0, 1)$. Then it is easy to show that

$$\lambda \leq \frac{(1 - \varepsilon)(1 + \varepsilon \alpha_{1,2})}{\alpha_{1,2}} \Longrightarrow \lambda \leq 1/\alpha_{1,2} - \varepsilon^2 \text{ and } \lambda - 1 \leq \frac{1 - \alpha_{1,2}\lambda}{\alpha_{1,2}\varepsilon}. \quad (5.1.16)$$

### 5.1.6 Convergence properties of FDRS

The paper [30] assumed the stepsize constraint $\gamma \in (0, 2\beta)$ in order to guarantee convergence of Algorithm 11. We now show that the parameter $\gamma$ can (possibly) be increased beyond $2\beta$, which can result in faster practical performance. The proof follows by constructing a new Lipschitz differentiable function $h$ so that the triple $(f, g, V)$ generates the same FDRS operator, $T_{FDRS}$, as $(f, g, V)$. This result was not included in [30].

**Lemma 5.1.1.** Define a function

$$h := g \circ P_V. \quad (5.1.17)$$

Then the FDRS operator associated to $(f, g, V)$ is identical to the FDRS operator associated to $(f, h, V)$. Let $1/\beta_V$ be the Lipschitz constant of $\nabla h$. Then $\beta_V \geq \beta$. In addition, let $\gamma \in (0, 2\beta_V)$. Then $T_{FDRS}$ is $\alpha_{FDRS}^V$-averaged where

$$\alpha_{FDRS}^V := \frac{2\beta_V}{4\beta_V - \gamma}. \quad (5.1.18)$$

**Proof.** The averaged property of $T_{FDRS}$ and the equivalence of FDRS operators follows from Part 7 of Proposition 5.1.1. The bound $\beta_V \geq \beta$ follows because for all $x, y \in \mathcal{H}$,

$$\|\nabla h(x) - \nabla h(y)\| = \|P_V \circ g \circ P_V(x) - P_V \circ g \circ P_V(y)\| \leq \|\nabla g \circ P_V(x) - \nabla g \circ P_V(y)\| \leq (1/\beta)\|P_V(x) - P_V(y)\| \leq (1/\beta)\|x - y\|$$

$\blacksquare$
There are cases where $\beta_V$ is significantly larger than $\beta$. For instance, in the quadratic programming example in (5.1.5), $\beta$ is the reciprocal of the Lipschitz constant of $Q$, which is the maximal eigenvalue $\lambda_{\text{max}}(Q)$ of $Q$. On the other hand, the gradient $\nabla h = P_V \circ Q \circ P_V$ has rank at most $d - \text{rank}(A)$. Thus, unless the eigenvectors of $Q$ with eigenvalue $\lambda_{\text{max}}(Q)$ lie in the $(d - \text{rank}(A))$-dimensional space $V$, the constant $\beta_V = 1/\lambda_{\text{max}}(P_V \circ Q \circ P_V)$ is larger than $\beta = 1/\lambda_{\text{max}}(Q)$.

Most of our results do not require that $(z^j)_{j \geq 0}$ converges. However, for completeness we include the following weak convergence result.

**Proposition 5.1.3.** Let $\gamma \in (0, 2\beta_V)$, let $(\lambda_j)_{j \geq 0} \subseteq (0, 1/\alpha_V^{\text{FDRS}})$, and suppose that $\sum_{i=0}^{\infty} (1 - \lambda_i \alpha_V^{\text{FDRS}})/\lambda_i \alpha_V^{\text{FDRS}} = \infty$. Then $(z^j)_{j \geq 0}$ (from Algorithm 11) weakly converges to a fixed-point of $T_{\text{FDRS}}$.

**Proof.** Apply [30, Proposition 3.1] with the new averaged parameter $\alpha_V^{\text{FDRS}}$. \qed

The following theorem recalls several results on convergence rates for the iteration of averaged operators from Chapter 2. In addition, we show that $(\lambda_j \|\nabla h(z^j) - \nabla h(z^*)\|^2)_{j \geq 0}$ is a summable sequence [30] whenever $(\lambda_j)_{j \geq 0}$ is chosen properly.

**Theorem 5.1.1.** Suppose that $(z^j)_{j \geq 0}$ is generated by Algorithm 11 with $\gamma \in (0, 2\beta_V)$ and $(\lambda_j)_{j \geq 0} \subseteq (0, 1/\alpha_V^{\text{FDRS}})$, and let $z^*$ be a fixed-point of $T_{\text{FDRS}}$. Then

1. Fejér monotonicity: the sequence $(\|z^j - z^*\|^2)_{j \geq 0}$ is nonincreasing. In addition, for all $z \in \mathcal{H}$ and $\lambda \in (0, 1/\alpha_V^{\text{FDRS}})$, we have $\|(T_{\text{FDRS}})\lambda z - z^*\| \leq \|z - z^*\|$.

2. Summable fixed-point residual: The sum is finite:

$$\sum_{i=0}^{\infty} \frac{1 - \lambda_i \alpha_V^{\text{FDRS}}}{\lambda_i \alpha_V^{\text{FDRS}}} \|z^{i+1} - z^i\|^2 \leq \|z^0 - z^*\|^2.$$

3. Convergence rates of fixed-point residual: For all $k \geq 0$, let $\tau_k := (1 - \lambda_k \alpha_V^{\text{FDRS}}) \lambda_k / \alpha_V^{\text{FDRS}}$. 

190
Suppose that $\tau := \inf_{j \geq 0} \tau_j > 0$. Then for $\lambda > 0$ and $k \geq 0$,

$$\| (T_{\text{FDRS}}) \lambda (z^k) - z^k \|^2 \leq \frac{\lambda^2 \| z^0 - z^* \|^2}{\tau (k + 1)}; \quad (5.1.19)$$

$$\| (T_{\text{FDRS}}) \lambda (z^k) - z^k \|^2 = o \left( \frac{1}{k + 1} \right). \quad (5.1.20)$$

4. Gradient summability: Let $\varepsilon \in (0, 1)$ and suppose that

$$(\lambda_j)_{j \geq 0} \subseteq \left( 0, \frac{(1 - \varepsilon)(1 + \varepsilon \alpha^V_{FDRS})}{\alpha^V_{FDRS}} \right). \quad (5.1.21)$$

Then the following gradient sum is finite:

$$\sum_{i=0}^{\infty} \lambda_i \| \nabla h(z^i) - \nabla h(z^*) \|^2 \leq \frac{\gamma (1 + \varepsilon)}{\varepsilon (2 \beta^V - \gamma)} \| z^0 - z^* \|^2. \quad (5.1.22)$$

Proof. Parts 1, 2, and 3 are a direct consequence of Theorem 2.3.1 of Chapter 2 applied to the $\alpha^V_{FDRS}$-averaged operator $T_{\text{FDRS}}$. Part 4 is a direct consequence of Proposition 5.1.2 applied to the $(1/2)$-averaged operator $T_1 = ((1/2)I_H + (1/2)\text{refl}_f \circ \text{refl}_{\chi_V})$ (see Part 6 of Proposition 5.1.1) and the $(\gamma/(2 \beta^V))$-averaged operator $T_2 = I - \gamma \nabla h$ (from the Baillon-Haddad Theorem [3] and [11, Proposition 4.33]).

We call the following term is called the fixed-point residual (FPR):

$$\| T_{\text{FDRS}} z^k - z^k \|^2 = \frac{1}{\lambda_k^2} \| z^{k+1} - z^k \|^2 \quad (5.1.23)$$

Remark 5.1.3. Note that the convergence rate proved for $\| T_{\text{FDRS}} z^k - z^k \|^2$ in (5.1.19) is sharp for the $T_{\text{FDRS}}$ operator by the results of Section 2.6.1.1 of Theorem 2.

## 5.2 Subgradients and fundamental inequalities

In this section, we prove several algebraic identities of the FDRS algorithm. In addition, we prove a relationship between the FPR and the objective error (Propositions 5.2.1 and 5.2.2).

In first-order optimization algorithms, we only have access to (sub)gradients and function values. Consequently, the FPR is usually the squared norm of a linear combination
of (sub)gradients of the objective functions. For example, the gradient descent algorithm
for a smooth function \( f \) generates a sequence of iterates by using forward gradient steps:
\[ z_{k+1} = z_k - \nabla f(z_k); \] the FPR is \( \|z_{k+1} - z_k\|^2 = \|\nabla f(z_k)\|^2 \).

In splitting algorithms, the FPR is more complex because the subgradients are generated via forward-gradient or proximal (backward) steps (see Part 1 of Proposition 5.1.1) at different points. Thus, unlike the gradient descent algorithm where the objective error \( f(z_k) - f(x^*) \leq \langle z_k - x^*, \nabla f(x_k) \rangle \) can be bounded with the subgradient inequality, splitting algorithms for two or more functions can only bound the objective error when some or all of the functions are evaluated at separate points — unless a Lipschitz assumption is imposed. In order to use this Lipschitz assumption, we enforce consensus among the variables, which is why the FPR rate is useful.

### 5.2.1 A subgradient representation of FDRS

\[
-\gamma \left( \nabla \chi_V(x_h) + \nabla h(x_h) + \nabla f(x_f) \right)
\]

\[
-\gamma \nabla \chi_V(x_h)
\]

\[
\gamma \nabla \chi_V(x_h)
\]

\[
\lambda(x_f - x_h)
\]

\[
T_{\text{FDRS}}(z)\quad (T_{\text{FDRS}}\lambda(z)
\]

**Figure 5.1:** A single FDRS iteration, from \( z \) to \( (T_{\text{FDRS}}\lambda(z) \) (see Lemma 5.2.1).

Both occurrences of \( \nabla \chi_V(x_h) \) represent the same subgradient \((1/\gamma) P_{V^\perp} z = (1/\gamma)(z - x_h) \in V^\perp\).

Figure 5.1 pictures one iteration of Algorithm 11: FDRS projects \( z \) onto \( V \) to get \( x_h = z - \gamma \nabla \chi_V(x_h) \). The reflection of \( z \) across \( V \) is \( x_h - \gamma \nabla \chi_V(x_h) = z - 2\gamma \nabla \chi_V(x_h) \). Then FDRS takes a forward-gradient with respect to \( \nabla h(x_h) \) from the reflected point.
and a proximal (backward) step with respect to $f$ to get $x_f$. Finally, we move from $x_f$ to $T_{\text{FDRS}}z$ by traveling along the positive subgradient $\gamma \tilde{\nabla} \chi_V(x_h)$.

The following lemma is proved in Appendix B.2.

**Lemma 5.2.1 (FDRS identities).** Let $z \in H$. Define points $x_h$ and $x_f$:

$$x_h := P_V z \quad \text{and} \quad x_f := \text{prox}_{\gamma f} \circ \text{refl}_{\chi_V} \circ (I_H - \gamma \nabla h)(z).$$

Then the identities hold

$$x_h := z - \gamma \tilde{\nabla} \chi_V(x_h) \quad \text{and} \quad x_f := x_h - \gamma \left( \tilde{\nabla} \chi_V(x_h) + \nabla h(x_h) + \tilde{\nabla} f(x_f) \right) \quad (5.2.1)$$

where $\tilde{\nabla} \chi_V(x_h) = (1/\gamma) P_{V \perp}(z)$ and $\tilde{\nabla} f(x_f)$ is uniquely defined by Part 1 of Proposition 5.1.1. In addition, each FDRS step has the following form:

$$(T_{\text{FDRS}})_{\lambda}(z) - z = -\lambda(x_f - x_h) = -\gamma \lambda \left( \tilde{\nabla} \chi_V(x_h) + \nabla h(x_h) + \tilde{\nabla} f(x_f) \right). \quad (5.2.2)$$

In particular, $T_{\text{FDRS}}(z) = x_f + \gamma \tilde{\nabla} \chi_V(x_h)$.

### 5.2.2 Optimality conditions of FDRS

The following lemma characterizes the zeros of $\partial f + \nabla h + \partial \chi_V$ in terms of the fixed-points of the FDRS operator. The intuition is the following: If $z^*$ is a fixed-point of $T_{\text{FDRS}}$, then the base of the rectangle in Figure 5.1 has length zero. Thus, $x^* := x^*_h = x^*_f$, and if we travel around the perimeter of the rectangle, we will start and begin at $z^*$. This argument shows that $\gamma \tilde{\nabla} f(x^*) + \gamma \nabla h(x^*) + \gamma \tilde{\nabla} \chi_V(x^*) = 0$, i.e., $x^* \in \text{zer}(\partial f + \nabla h + \partial \chi_V)$.

The following lemma is proved in Appendix B.3.

**Lemma 5.2.2 (FDRS optimality conditions).** The following set equality holds:

$$\text{zer}(\partial f + \nabla h + \partial \chi_V) = \{ P_V z \mid z \in H, T_{\text{FDRS}}z = z \}$$

That is, if $z^*$ is a fixed-point of $T_{\text{FDRS}}$, then $x^* := P_V z^* = x^*_h = x^*_f$ is a minimizer of (5.1.2), and $z^* - x^* = P_V(z^*) = \gamma \tilde{\nabla} \chi_V(x^*_h) \in \partial \chi_V(x^*)$.  

193
5.2.3 Fundamental inequalities

In this section, we prove two fundamental inequalities that relate the FPR (see (5.1.23)) to the objective error.

Throughout the rest of the chapter, we use the following notation: The functions $f$ and $g$ are $\mu_f$ and $\mu_g$-strongly convex, respectively, where we allow $\mu_f$ or $\mu_g$ to be zero (i.e., no strong convexity). In addition, we assume that $f$ is $(1/\beta_f)$-Lipschitz differentiable, where we allow $\beta_f = 0$. If $\beta_f > 0$, then $\tilde{\nabla} f = \nabla f$. With these assumptions, we get the following lower bounds [11, Theorem 18.15]:

$$\forall x, y \in \text{dom}(f) \quad f(x) \geq f(y) + \langle x - y, \tilde{\nabla} f(y) \rangle + S_f(x, y), \quad (5.2.3)$$

$$\forall x, y \in \mathcal{H} \quad h(x) \geq h(y) + \langle x - y, \nabla h(y) \rangle + S_h(x, y), \quad (5.2.4)$$

where for all $\tilde{\nabla} f(x) \in \partial f(x)$ and $\tilde{\nabla} f(y) \in \partial f(y)$, we have

$$S_f(x, y) := \max \left\{ \frac{\mu_f}{2} \|x - y\|^2, \frac{\beta_f}{2} \|\tilde{\nabla} f(x) - \tilde{\nabla} f(y)\|^2 \right\}; \quad (5.2.5)$$

$$S_h(x, y) := \max \left\{ \frac{\mu_g}{2} \|P_V x - P_V y\|^2, \frac{\beta_V}{2} \|\nabla h(x) - \nabla h(y)\|^2 \right\}. \quad (5.2.6)$$

Note that $S_f(x, y) \geq 0$ and $S_h(x, y) \geq 0$ for any $x, y \in \text{dom}(f)$.

See Appendices B.4, B.5, and B.6 for the proofs of the following inequalities:

**Proposition 5.2.1** (Upper fundamental inequality). Let $z \in \mathcal{H}$, let $\lambda > 0$, and let $z^+ = (T_{FDRS})_\lambda(z)$. Then for all $x \in V \cap \text{dom}(f)$, we have the following inequality:

$$2\gamma \lambda (f(x_f) + h(x_h) - f(x) - h(x) + S_f(x_f, x) + S_h(x_h, x))$$

$$\leq \|z - x\|^2 - \|z^+ - x\|^2 + \left(1 - \frac{2}{\lambda}\right) \|z^+ - z\|^2 + 2\gamma \langle \nabla h(x_h), z - z^+ \rangle \quad (5.2.7)$$

where $x_f$ and $x_h$ are defined as in Lemma 5.2.1.

**Proposition 5.2.2** (Lower fundamental inequality). Let $z^* \in \mathcal{H}$ be a fixed-point of $T_{FDRS}$, and let $x^* := P_V z^*$. Choose subgradients $\tilde{\nabla} f(x^*) \in \partial f(x^*)$ and $\tilde{\nabla} \chi_V(x^*) \in \partial \chi_V(x^*)$ with
\[ \tilde{\nabla} f(x^*) + \nabla h(x^*) + \tilde{\nabla} \chi_V(x^*) = 0 \quad \text{(see Lemma 5.2.2).} \]

Then for all \( x_f \in \text{dom}(f) \) and \( x_h \in V \), we have

\[
f(x_f) + h(x_h) - f(x^*) - g(x^*) \geq \langle x_f - x_h, \tilde{\nabla} f(x^*) \rangle + S_f(x_f, x^*) + S_h(x_h, x^*). \tag{5.2.8}
\]

**Corollary 5.2.1.** Let \( z \in \mathcal{H} \), let \( \lambda > 0 \), and let \( z^+ = (T_{\text{FDRS}})_\lambda(z) \). Let \( z^* \in \mathcal{H} \) be a fixed-point of \( T_{\text{FDRS}} \), and let \( x^* := P_V z^* \). Then with \( x_f \) and \( x_h \) from Lemma 5.2.1,

\[
4\gamma \lambda (S_f(x_f, x^*) + S_h(x_h, x^*)) \leq \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left( 1 - \frac{2}{\lambda} \right) \|z^+ - z\|^2
\]

\[ + 2\gamma \langle \nabla h(x_h) - \nabla h(x^*), z - z^+ \rangle. \tag{5.2.9}\]

### 5.3 Objective convergence rates

In this section, we analyze the ergodic and nonergodic convergence rates of the FDRS algorithm applied to (5.1.2).

Throughout the rest of the chapter, \( z^* \) will denote an arbitrary fixed-point of \( T_{\text{FDRS}} \), and we define a minimizer of (5.1.2) using Lemma 5.2.2: \( x^* := P_V z^* \).

All of our bounds will be produced on objective errors of the form:

\[
f(x_f^k) + h(x_h^k) - f(x^*) - g(x^*) \quad \text{and} \quad f(x_f^k) + h(x_h^k) - f(x^*) - g(x^*). \tag{5.3.1}
\]

The objective error on the left hand side of (5.3.1) can be negative. Thus, we bound its absolute value. In addition, we bound \( \|x_f^k - x_f^k\| \). Because \( x_h^k \in V \), the objective error on the right hand size of (5.3.1) is positive. Consequently, \( x_h^k \) is the natural point at which to measure the convergence rate. To derive such a bound, we assume \( f \) is Lipschitz. Note that in both cases, we have the identity \( h(x_h^k) = (g \circ P_V)(x_h^k) = g(x_h^k) \).

Finally, all of our lower bounds will involve the subgradient norms \( \|\tilde{\nabla} f(x^*)\|, \|\nabla h(x^*)\|, \) and \( \|\tilde{\nabla} \chi_V(x^*)\| \). We can always assume these norms to be minimal over all \( \tilde{\nabla} f(x^*) \) satisfying \( \tilde{\nabla} f(x^*) + \nabla h(x^*) + \tilde{\nabla} \chi_V(x^*) = 0 \) (See Proposition 5.2.1). We make this assumption throughout the rest of the chapter.
5.3.1 Ergodic convergence rates

In this section, we analyze the ergodic convergence rate of the FDRS algorithm. The key idea is to use the telescoping property of the upper and lower fundamental inequalities, together with the summability of the difference of gradients shown in Part 4 of Theorem 5.1.1. See Section 5.1.2 for the distinction between ergodic and nonergodic convergence rates.

**Theorem 5.3.1** (Ergodic convergence of FDRS). Let \( \gamma \in (0, 2\beta_V) \), let \( \varepsilon \in (0, 1) \), and suppose that \((\lambda_j)_{j \geq 0}\) satisfies (5.1.21). Define \((\bar{x}_j^f)_{j \geq 0}\) and \((\bar{x}_j^g)_{j \geq 0}\) as in (5.1.7). Then we have the following convergence rate: for all \( k \geq 0 \),

\[
-2\|z^0 - z^*\|\left\|\widetilde{\nabla} f(x^*)\right\| \leq f(\bar{x}_k^f) + h(\bar{x}_k^h) - f(x^*) - h(x^*)
\]

\[
\leq \left( \|z^0 - z^*\| + 4\gamma\|\nabla h(x^*)\| + \frac{(1+\varepsilon)\gamma\|z^0 - z^*\|}{\varepsilon^3(2\beta_V - \gamma)} \right) \|z^0 - z^*\|.
\]

In addition the following feasibility bound holds: \( \|\bar{x}_k^f - \bar{x}_k^h\| \leq (2/\Lambda_k)\|z^0 - z^*\| \).

**Proof.** Fix \( k \geq 0 \). The feasibility bound follows from Part 1 of Proposition 1:

\[
\|\bar{x}_k^f - \bar{x}_k^h\| = \left\| \frac{1}{\Lambda_k} \sum_{i=0}^{k} (z^{i+1} - z^i) \right\| = \frac{1}{\Lambda_k} \|z^0 - z^{k+1}\| \leq \frac{1}{\Lambda_k} \left( \|z^0 - z^*\| + \|z^* - z^{k+1}\| \right)
\]

\[
\leq \frac{2}{\Lambda_k} \|z^0 - z^*\|. \tag{5.3.2}
\]

Now we prove the objective convergence rates. For all \( k \geq 0 \), let \( \eta_k := 2/\lambda_k - 1 \). Note that \( \eta_k > 0 \) by (5.1.16) because we have \( \lambda_k < 1/\alpha'_{FDRS} - \varepsilon^2 \leq 2 - \varepsilon^2 \) and \( 1/\eta_k = \lambda_k/(2 - \lambda_k) \leq \lambda_k/\varepsilon^2 \). Thus, by Cauchy-Schwarz and (5.1.12), we have

\[
2\gamma\langle \nabla h(x_k^h), z^k - z^{k+1} \rangle = 2\gamma\langle \nabla h(x^*), z^k - z^{k+1} \rangle + 2\gamma\langle \nabla h(x_k^h) - \nabla h(x^*), z^k - z^{k+1} \rangle
\]

\[
\leq 2\gamma\langle \nabla h(x^*), z^k - z^{k+1} \rangle + \frac{\gamma^2}{\eta_k} \|\nabla h(x_k^h) - \nabla h(x^*)\|^2 + \eta_k \|z^k - z^{k+1}\|^2. \tag{5.3.3}
\]

Therefore, by Jensen’s inequality, the Cauchy-Schwarz inequality, (5.2.7), and the bound
\[ \|z^0 - z^{k+1}\| \leq 2\|z^0 - z^*\| \text{ (see (5.3.2))}, \]

we have

\[
f(x_f^k) + h(x_h^k) - f(x^*) - h(x^*) \leq \frac{1}{\Lambda_k} \sum_{i=0}^{k} \lambda_i \left( f(x_f^i) + h(x_h^i) - f(x^*) - h(x^*) \right)
\]

(5.2.7)

\[
\leq \frac{1}{2\gamma \Lambda_k} \sum_{i=0}^{k} \left( \|z^i - x^*\|^2 - \|z^{i+1} - x^*\|^2 - \eta_i \|z^{i+1} - z^i\|^2 + 2\gamma \langle \nabla h(x_h^i), z^i - z^{i+1} \rangle \right)
\]

(5.3.3)

\[
\leq \frac{1}{2\gamma \Lambda_k} \left( \|z^0 - x^*\|^2 + 2\gamma \langle \nabla h(x^*), z^0 - z^{k+1} \rangle + (\gamma^2/\varepsilon^2) \sum_{i=0}^{\infty} \lambda_i \|\nabla h(x_h^i) - \nabla h(x^*)\|^2 \right)
\]

(5.1.22)

\[
\leq \frac{2\gamma \Lambda_k}{\|z^0 - x^*\|^2 + 4\gamma \|\nabla h(x^*)\| \|z^0 - z^*\| + (1 + \varepsilon)\gamma^3 \|z^0 - z^*\|^2 /(\varepsilon^3(2\beta_V - \gamma))}.
\]

The lower bound in Proposition 5.2.2 and the Cauchy-Schwarz inequality show that

\[
f(x_f^k) + h(x_h^k) - f(x^*) - h(x^*) \geq \langle x_f^k - x_h^k, \tilde{\nabla} f(x^*) \rangle \geq -\|x_f^k - x_h^k\| \|\tilde{\nabla} f(x^*)\|
\]

\[
\geq -\frac{2\|z^0 - z^*\| \|\tilde{\nabla} f(x^*)\|}{\Lambda_k}.
\]

\[
\square
\]

In general, \(x_f^k\) and \(x_h^k\) are not in \(\text{dom}(f)\). However, the conclusion of Theorem 5.3.1 can be improved if \(f\) is Lipschitz continuous. The following proposition gives a sufficient condition for Lipschitz continuity on a ball.

**Proposition 5.3.1 (Lipschitz continuity on a ball [11, Proposition 8.28])**. Suppose that \(f : H \to (-\infty, \infty]\) is proper and convex. Let \(\rho > 0\), and let \(x_0 \in \text{dom}(f)\). If \(\delta = \sup_{x,y \in B(x_0, 2\rho)} |f(x) - f(y)| < \infty\), then \(f\) is \((\delta/\rho)\)-Lipschitz on \(B(x_0, \rho)\).

To use this fact, we need to show that the sequences \((x_f^j)_{j \geq 0}\), and \((x_h^j)_{j \geq 0}\) are bounded. Recall that \(x_f^s = \text{prox}_{\gamma_f} \circ \text{refl}_{x_V} \circ (I_{H} - \gamma \nabla h)(x^s)\) for \(s \in \{*, k\}\). Proximal, reflection, and forward-gradient maps are nonexpansive (see Proposition 5.1.1, the Baillon-Haddad Theorem [3], and [11, Proposition 4.33]), so we have \(\max\{\|x_f^k - x^*\|, \|x_h^k - x^*\|\} \leq \|z^k - z^*\| \leq \|z^0 - z^*\|\). Thus, \((x_f^j)_{j \geq 0}, (x_h^j)_{j \geq 0} \subseteq B(x^*, \|z^0 - z^*\|)\). The ball is convex, so \((\pi_f^j)_{j \geq 0}, (\pi_h^j)_{j \geq 0} \subseteq B(x^*, \|z^0 - z^*\|)\).
Corollary 5.3.1 (Ergodic convergence with Lipschitz f). Let the notation be as in Theorem 5.3.1. Let \( L \geq 0 \) and suppose \( f \) is \( L \)-Lipschitz on \( B(x^*, \|z^0 - z^*\|) \). Then

\[
0 \leq f(x^k_h) + h(x^k_h) - f(x^*) - h(x^*) \leq \left( \|z^0 - z^*\| + 2\gamma \|\nabla h(x^*)\| + \frac{(1+\varepsilon)\gamma\|z^0 - z^*\|}{\varepsilon(2\beta_V - \gamma)} \right) \|z^0 - z^*\| + \frac{2L\|z^0 - z^*\|}{\Lambda_k}.
\]

Proof. The proof follows from by combining the upper bound in Theorem 5.3.1 with the following bound: \( f(x^k_h) \leq f(x^k_f) + L\|x^k_f - x^k_h\| \leq f(x^k_f) + 2L\|z^0 - z^*\|/\Lambda_k \).

Remark 5.3.1. Corollary 5.3.1 is sharp by Proposition 2.7.2 of Chapter 2.

5.3.2 Nonergodic convergence rates

In this section, we analyze the nonergodic convergence rate of FDRS when \( (\lambda_j)_{j \geq 0} \) is bounded away from 0 and \( 1/\alpha^V_{\text{FDRS}} \). The proof bounds the inequalities in Propositions 5.2.1 and 5.2.2 with Theorem 5.1.1.

Theorem 5.3.2 (Nonergodic convergence of FDRS). For all \( k \geq 0 \), let \( \lambda_k \in (0,1/\alpha^V_{\text{FDRS}}) \). Suppose that \( \tau := \inf_{j \geq 0} (1 - \alpha^V_{\text{FDRS}})\lambda_j / \alpha^V_{\text{FDRS}} > 0 \). Then

\[
\|x^k_f - x^k_h\| \leq \frac{\|z^0 - z^*\|}{\sqrt{\tau(k+1)}}, \quad \|x^k_f - x^k_h\| = o\left(\frac{1}{\sqrt{k+1}}\right),
\]

and

\[
-\frac{\|z^0 - z^*\| \|
abla f(x^*)\|}{\sqrt{\tau(k+1)}} \leq f(x^k_f) + h(x^k_h) - f(x^*) - g(x^*) \leq \frac{\|z^* - x^*\| + (1 + \gamma/\beta_V)\|z^0 - z^*\| + \gamma\|\nabla h(x^*)\|}{\gamma\sqrt{\tau(k+1)}} \|z^0 - z^*\|,
\]

and \( |f(x^k_f) + h(x^k_h) - f(x^*) - g(x^*)| = o(1/\sqrt{k+1}) \).

Proof. First we note that \( \|\nabla h(x^k_h)\|_{j \geq 0} \) is bounded: for all \( k \geq 0 \),

\[
\|\nabla h(x^k_h)\| \leq \|\nabla h(x^k_h) - \nabla h(x^*)\| + \|\nabla h(x^*)\| = \|\nabla h(z^k) - \nabla h(z^*)\| + \|\nabla h(x^*)\| \leq \frac{1}{\beta_V} \|z^k - z^*\| + \|\nabla h(x^*)\| \leq \frac{1}{\beta_V} \|z^0 - z^*\| + \|\nabla h(x^*)\|.
\]
because \( \| z^j - z^* \|_{j \geq 0} \) is decreasing (see Part 1 of Theorem 5.1.1).

Next fix \( k \geq 0 \). For any \( \lambda > 0 \), define \( \lambda = (T_{\text{FDRS}})_{\lambda}(z^k) \). Observe that \( x^k_f \) and \( x^k_h \) do not depend on the value of \( \lambda_k \). Therefore, by Proposition 5.2.1 and Lemma 5.2.1,

\[
f(x^k_f) + h(x^k_h) - f(x^*) - g(x^*) \leq \inf_{\lambda \in [0,1/\alpha_{\text{FDRS}}^V]} \frac{1}{2\gamma \lambda} \left( \| z^k - x^* \|^2 - \| z_\lambda - x^* \|^2 + \left( 1 - \frac{2}{\lambda} \right) \| z_\lambda - z^k \|^2 \right.
\]

\[
+ 2\gamma \langle \nabla h(x^k_h), z^k - z_\lambda \rangle
\]

\[(5.1.11) \leq \inf_{\lambda \in [0,1/\alpha_{\text{FDRS}}^V]} \frac{1}{2\gamma \lambda} \left( 2\| z_\lambda - x^*, z^k - z_\lambda \| + 2 \left( 1 - \frac{1}{\lambda} \right) \| z_\lambda - z^k \|^2 \right.
\]

\[
+ 2\gamma \langle \nabla h(x^k_h), z^k - z_\lambda \rangle
\]

\[(5.3.4) \leq \frac{1}{2\gamma} \left( 2\| z_1 - x^*, z^k - z_1 \| + 2\gamma \left( \frac{1}{\beta^V} \| z^0 - z^* \| + \| \nabla h(x^*) \| \right) \| z_1 - z^k \| \right)
\]

\[(5.1.19) \leq \left( \| z_1 - x^* \| + \| \nabla h(x^*) \| \right) \left( \| z^0 - z^* \| \right)
\]

where we use \( \| z_1 - x^* \| \leq \| z_1 - z^* \| + \| z^* - x^* \| \leq \| z^0 - z^* \| + \| z^* - x^* \| \) (Theorem 5.1.1).

The lower bound follows from (5.2.8) and Part 3 of Theorem 5.1.1:

\[
f(x^k_f) + h(x^k_h) - f(x^*) - g(x^*) \geq \langle x^k_f - x^k_h, \nabla f(x^*) \rangle = \frac{1}{\lambda_k} \langle z^{k+1} - z^k, \nabla f(x^*) \rangle
\]

\[(5.1.19) \geq \frac{\| z^0 - z^* \| \| \nabla f(x^*) \|}{\sqrt{k+1}}.
\]

The \( o(1/\sqrt{k+1}) \) rates follow from (5.3.5) and (5.3.6), and the corresponding rates for the FPR in (5.1.19). The bounds on \( x^k_f - x^k_h \) follow from \( x^k_f - x^k_h = T_{\text{FDRS}} z^k - z^k \).

If \( f \) is Lipschitz continuous, we can evaluate the entire objective function at \( x^k_h \). The proof of the following corollary is analogous to Corollary 5.3.1. We ask the reader to recall from Section 5.3.1 that \( (x^j_f)_{j \geq 0}, (x^j_h)_{j \geq 0} \subseteq B(x^*, \| z^0 - z^* \|) \).
Corollary 5.3.2 (Nonergodic convergence with Lipschitz f). Let the notation be as in Theorem 5.3.2. Let \( L \geq 0 \) and suppose \( f \) is \( L \)-Lipschitz on \( B(x^*, \|z^0 - z^*\|) \). Then

\[
0 \leq f(x^k_h) + h(x^k_h) - f(x^*) - h(x^*) - \frac{\|z^* - x^*\| + (1 + \gamma/\beta_V)\|z^0 - z^*\| + \gamma\|\nabla h(x^*)\|}{\gamma\sqrt{\tau(k + 1)}} \|z^0 - z^*\| + \frac{L\|z^0 - z^*\|}{\sqrt{\tau(k + 1)}},
\]

and \( f(x^k_h) + h(x^k_h) - f(x^*) - h(x^*) = o(1/\sqrt{k + 1}) \).

Proof. Combine the upper bound in Theorem 5.3.2 with the following bound: \( f(x^k_h) \leq f(x^k_f) + L\|x^k_f - x^k_h\| \leq f(x^k_f) + L\|z^0 - z^*\|/\sqrt{\tau(k + 1)} \). The \( o(1/\sqrt{k + 1}) \) rate follows because \( \|x^k_f - x^k_h\| = \|T_{FDRS}^k z^k - z^k\| = o(1/\sqrt{k + 1}) \) (see (5.2.2) and (5.1.19)) and \( |f(x^k_f) + h(x^k_h) - f(x^*) - h(x^*)| = o(1/\sqrt{k + 1}) \) (see Theorem 5.3.2).

Remark 5.3.2. Corollary 5.3.2 is sharp by Theorem 2.7.2 of Chapter 2.

5.4 Strong convexity

In this section, we show that \( (x^j_f)_{j \geq 0}, (x^j_h)_{j \geq 0} \), and their ergodic variants converge strongly whenever \( f \) or \( g \) is strongly convex. The techniques in this section are similar to those in Section 5.3, so we defer the proof to Appendix B.7

Theorem 5.4.1 (Auxiliary term bound). Let \( \gamma \in (0, 2\beta_V) \), let \( (\lambda_j)_{j \geq 0} \subseteq (0, 1/O_{FDRS}) \), let \( z^0 \in \mathcal{H} \), and suppose that \( (z^j)_{j \geq 0} \) is generated by Algorithm 11. Then

1. “Best” iterate convergence: Let \( \varepsilon \in (0, 1) \) and suppose that \( (\lambda_j)_{j \geq 0} \) satisfies (5.1.21).
   If \( \Delta := \inf_{j \geq 0} \lambda_j > 0 \), then
   \[
   \min_{0 \leq j \leq k} \left( S_f(x^j_f, x^*) + S_h(x^j_h, x^*) \right) \leq \frac{\left(1 + \frac{(1+\varepsilon)^3}{\varepsilon^3(2\beta_V - \gamma)}\right)\|z^0 - z^*\|^2}{4\gamma\Delta(k + 1)}.
   \]
   and, thus, \( \min_{0 \leq j \leq k} S_f(x^j_f, x^*) = o(1/(k + 1)) \) and \( \min_{0 \leq j \leq k} S_h(x^j_h, x^*) = o(1/(k + 1)) \).

2. Ergodic convergence: If \( \varepsilon \in (0, 1) \), and \( (\lambda_j)_{j \geq 0} \) satisfies (5.1.21), then
   \[
   \overline{S}_f^k + \overline{S}_h^k \leq \frac{\left(1 + \frac{(1+\varepsilon)^3}{\varepsilon^3(2\beta_V - \gamma)}\right)\|z^0 - z^*\|^2}{4\gamma\Lambda_k}.
   \]

200
where \( S_f^k := \max \left\{ \frac{\mu}{2} \| x_f^k - x^* \|^2, \frac{\beta_f}{2} \left\| \lambda \sum_{i=0}^{k} \lambda_i \left( \tilde{\nabla} f(x_f^k) - \tilde{\nabla} f(x^*) \right) \right\|^2 \right\} \) and \( S_h^k \) is similarly defined.

3. Nonergodic convergence: If \( \tau := \inf_{j \geq 0} \left( 1 - \alpha V_{FDRS} \lambda_j / \alpha_{FDRS} \right) > 0 \), then \( S_f(x_f^k, x^*) + S_h(x_h^k, x^*) = o(1/\sqrt{k+1}) \) and

\[
S_f(x_f^k, x^*) + S_h(x_h^k, x^*) \leq \frac{(1 + \gamma/\beta_V)\| z^0 - z^* \|^2}{\gamma\sqrt{\tau(k+1)}},
\]

**Remark 5.4.1.** See Section 5.6.1 for a proof that the nonergodic “best” rates are sharp. It is not clear if we can improve the general nonergodic rates to \( o(1/(k+1)) \).

### 5.5 Lipschitz differentiability

In this section, we assume \( f \) is smooth:

**Assumption 5.5.1.** \( f \) is differentiable and \( \nabla f \) is \((1/\beta_f)\)-Lipschitz where \( \beta_f > 0 \).

Under Assumption 5.5.1, we will show that the objective value

\[
f(x_h^k) + h(x_h^k) - f(x^*) - h(x^*) = f(x_h^k) + g(x_h^k) - f(x^*) - g(x^*)
\]

is summable. Therefore, by Lemma 2.2.1 of Chapter 2 the minimal objective error after \( k \) iterations is of order \( o(1/(k+1)) \). We will need the following upper bound to prove this. See Appendix B.8 for the proof.

**Proposition 5.5.1** (Fundamental inequality under Assumption 5.5.1). If \( \gamma \in (0, 2\beta_V) \), \( \lambda > 0 \), \( z \in \mathcal{H} \), \( z^+ = (T_{FDRS})_\lambda(z) \), \( z^* \) is a fixed-point of \( T_{FDRS} \), and \( x^* = P_V z^* \), then

\[
2\gamma \lambda (f(x_h) + h(x_h) - f(x^*) - g(x^*)) \leq \begin{cases} 
\| z - z^* \|^2 - \| z^+ - z^* \|^2 + \left( 1 + \frac{\gamma - \beta_f}{\beta_f \lambda} \right) \| z - z^+ \|^2 \\
+2\gamma \langle \nabla h(x_h) - \nabla h(x^*), z - z^+ \rangle & \text{if } \gamma \leq \beta_f \\
\left( 1 + \frac{\gamma - \beta_f}{2\beta_f} \right) \left( \| z - z^* \|^2 - \| z^+ - z^* \|^2 + \| z - z^+ \|^2 \right) \\
+2\gamma \left( 1 + \frac{\gamma - \beta_f}{2\beta_f} \right) \langle \nabla h(x_h) - \nabla h(x^*), z - z^+ \rangle & \text{if } \gamma > \beta_f.
\end{cases}
\]
The next theorem shows that the upper bound in Proposition 5.5.1 is summable and, as a consequence, we will have \( o(1/(k+1)) \) convergence.

**Theorem 5.5.1** (Convergence rates under Assumption 5.5.1). Let \( \gamma \in (0, 2\beta_V) \), let \( \varepsilon \in (0, 1) \), and suppose \((\lambda_j)_{j \geq 0}\) satisfies (5.1.21). Suppose that \( \tau := \inf_{j \geq 0} (1 - \alpha FDRS/\lambda_j) \beta_V \lambda_j > 0 \) and let \( \Delta = \inf_{j \geq 0} \lambda_j > 0 \). Let \( z^0 \in \mathcal{H} \), let \( z^* \) be a fixed-point of \( T_{FDRS} \), and let \( x^* = P_V z^* \).

Then
\[
\min_{0 \leq j \leq k} \left( f(x_h^j) + h(x_h^j) - f(x^*) - h(x^*) \right) = o \left( \frac{1}{k+1} \right).
\]

**Proof.** First recall that, by Part 2 of Theorem 5.1.1, we have
\[
\sum_{i=0}^{\infty} \| z^{i+1} - z^i \|^2 \leq \frac{1}{\tau} \sum_{i=0}^{\infty} \frac{1 - \lambda_i \alpha_{FDRS}}{\lambda_i \alpha_{FDRS}} \| z^{i+1} - z^i \|^2 \leq \frac{1}{\tau} \| z^0 - z^* \|^2.
\]

Next, we use the Cauchy-Schwarz inequality and (5.1.12) to show that
\[
\sum_{i=0}^{\infty} 2\gamma (\nabla h(x_h^i) - \nabla h(x^*), z^i - z^{i+1}) \leq \sum_{i=0}^{\infty} \left( \lambda_i \gamma^2 \| \nabla h(x_h^i) - \nabla h(x^*) \|^2 + \frac{1}{\lambda_i} \| z^i - z^{i+1} \|^2 \right) \leq \frac{(1 + \varepsilon) \gamma^3}{\varepsilon (2\beta_V - \gamma)} + \frac{1}{\lambda \tau} \| z^0 - z^* \|^2.
\]

If we combine the previous two sum bounds with (5.5.1), we get
\[
\sum_{i=0}^{\infty} (f(x_h^i) + h(x_h^i) - f(x^*) - h(x^*)) \leq \left( 1 + \frac{1}{\tau} + \frac{(1+\varepsilon) \gamma^3}{\varepsilon (2\beta_V - \gamma)} + \frac{1}{\lambda \tau} \right) \| z^0 - z^* \|^2 \times \begin{cases} 1 & \text{if } \gamma \leq \beta_f; \\ \left( 1 + \frac{\gamma - \beta_f}{2\beta_f} \right)^2 & \text{if } \gamma > \beta_f. \end{cases}
\]

The convergence rate now follows from Lemma 2.2.1 of Chapter 2.

**Remark 5.5.1.** Theorem 5.5.1 is sharp under Assumption 5.5.1 by Theorem 2.7.3 of Chapter 2.
5.6 Linear convergence

In this section, we prove FDRS converges linearly when \( \beta_f(\mu_g + \mu_f) > 0 \).

**Theorem 5.6.1 (Linear convergence).** Let \( \gamma \in (0, 2\beta_V) \), let \( (\lambda_j)_{j \geq 0} \subseteq (0, 1/\alpha_{FDRS}^V) \), let \( z^0 \in \mathcal{H} \), let \( z^* \) be a fixed-point of \( T_{FDRS} \), and let \( x^* = P_V z^* \). Let \( c > 1/2 \), let \( \gamma < \beta_V/c \), and let \( (\lambda_j)_{j \geq 0} \subseteq (0, (2c - 1)/c) \). For all \( \lambda \in (0, (2c - 1)/c) \), define

\[
C_1(\lambda) = \left( 1 - \frac{\lambda}{3} \min \left\{ \frac{\gamma \mu_g}{(1 + \gamma/\beta_V)^2}, \frac{\beta_f}{\gamma} \left( \frac{2c - 1}{c} - \lambda \right) \right\} \right)^{1/2};
\]

\[
C_2(\lambda) = \left( 1 - \frac{\lambda}{3} \min \left\{ \frac{\gamma \mu_f}{(1 + \gamma/\beta_f)^2}, \frac{\beta_V}{\gamma} \left( \frac{2c - 1}{c} - \lambda \right) \right\} \right)^{1/2}.
\]

Then for all \( k \geq 0 \), we have

\[
\|z^{k+1} - z^*\| \leq \|z^k - z^*\| \times \begin{cases} C_1(\lambda_k) & \text{if } \mu_g \beta_f > 0; \\ C_2(\lambda_k) & \text{if } \mu_f \beta_f > 0; \end{cases}
\]

\[
\|z^{k+1} - z^*\| \leq \|z^0 - z^*\| \times \begin{cases} \prod_{i=0}^k C_1(\lambda_i) & \text{if } \mu_g \beta_f > 0; \\ \prod_{i=0}^k C_2(\lambda_i) & \text{if } \mu_f \beta_f > 0. \end{cases}
\]

**Proof.** (5.2.9) shows that for all \( k \geq 0 \), we have

\[
\gamma \lambda_k \mu_f \|x^k_f - x^*\|^2 + \gamma \lambda_k \beta_f \|\nabla f(x^k_f) - \nabla f(x^*)\|^2 \\
+ \gamma \lambda_k \mu_g \|x^k_h - x^*\|^2 + \gamma \lambda_k \beta_V \|\nabla h(x^k_h) - \nabla h(x^*)\|^2 \\
\leq \|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2 + \left( 1 - \frac{2}{\lambda_k} \right) \|z^{k+1} - z^k\|^2 \\
+ 2\gamma \langle \nabla h(x^k_h) - \nabla h(x^*), z^k - z^{k+1} \rangle.
\]

In addition, by the Cauchy-Schwarz inequality and (5.1.12), we have

\[
2\gamma \langle \nabla h(x^k_h) - \nabla h(x^*), z^k - z^{k+1} \rangle \leq c\gamma^2 \lambda_k \|\nabla h(x^k_h) - \nabla h(x^*)\|^2 + \frac{1}{c\lambda_k} \|z^k - z^{k+1}\|^2.
\]

203
Therefore, for all \( k \geq 0 \),

\[
\gamma \lambda_k \mu_f \|x_f^k - x^*\|^2 + \gamma \lambda_k \beta_f \|\nabla f(x_f^k) - \nabla f(x^*)\|^2 \\
+ \gamma \lambda_k \mu_g \|x_h^k - x^*\|^2 + \gamma \lambda_k (\beta_V - c\gamma) \|\nabla h(x_h^k) - \nabla h(x^*)\|^2 \\
\leq \|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2 + \left(1 - \frac{2c - 1}{c\lambda_k}\right) \|z^{k+1} - z^k\|^2.
\]

Recall that we assume \( 1 - (2c - 1)/(c\lambda_k) < 0 \) and \( \beta_V - c\gamma > 0 \).

Now suppose that \( \beta_f \mu_g > 0 \). The following identity follows from from Lemma 5.2.1:

\[
z^k = T_{\text{FDRS}}(z^k) + (z^k - T_{\text{FDRS}}(z^k)) = x_h^k - \gamma \nabla h(x_h^k) - \gamma \nabla f(x_f^k) + \frac{1}{\lambda_k} (z^k - z^{k+1}).
\]

This identity results from tracing the perimeter of Figure 5.1 from \( x_h \) to \( x_f \) to \( T_{\text{FDRS}}z^k \) to \( z^k \). Likewise, we have \( z^* = x^* - \gamma \nabla h(x^*) - \gamma \nabla f(x^*) \).

Note that

\[
\|(x_h^k - \gamma \nabla h(x_h^k)) - (x^* - \gamma \nabla h(x^*))\| \leq \|x_h^k - x^*\| + \|\nabla h(x_h^k) - \nabla h(x^*)\| \\
\leq (1 + \gamma/\beta_V) \|x_h^k - x^*\|. \tag{5.6.2}
\]

Now, let \( C'_1 = 3 \max\left\{ (1 + \gamma/\beta_V)^2/(\gamma \lambda_k \mu_g), \gamma^2/(\gamma \lambda_k \beta_f), (1/\lambda_k^2) \left(\frac{2c - 1}{c\lambda_k} - 1\right)^{-1}\right\} \). By the convexity of \( \| \cdot \|^2 \), we have

\[
\|z^k - z^*\|^2 \leq 3(1 + \gamma/\beta_V)^2 \|x_h^k - x^*\|^2 + 3\gamma^2 \|\nabla f(x_f^k) - \nabla f(x^*)\|^2 + \frac{3}{\lambda_k^2} \|z^{k+1} - z^k\|^2 \\
\leq C'_1 \left( \gamma \lambda_k \mu_g \|x_h^k - x^*\|^2 + \gamma \lambda_k \beta_f \|\nabla f(x_f^k) - \nabla f(x^*)\|^2 + \left(\frac{2c - 1}{c\lambda_k} - 1\right) \|z^{k+1} - z^k\|^2 \right) \\
\leq C'_1 \|z^k - z^*\|^2 - C'_1 \|z^{k+1} - z^*\|^2.
\]

Therefore, \( \|z^{k+1} - z^*\| \leq (1 - (1/C'_1))^{1/2} \|z^k - z^*\| \).

Now assume that \( \beta_f \mu_f > 0 \). Observe that:

\[
z^k = x_h^k - \gamma \nabla h(x_h^k) - \gamma \nabla f(x_f^k) + \frac{1}{\lambda_k} (z^k - z^{k+1}) \\
= x_f^k - \gamma \nabla h(x_h^k) - \gamma \nabla f(x_f^k) + \frac{2}{\lambda_k} (z^k - z^{k+1})
\]

204
where we use the identity $x_k^h - x_j^f = (1/\lambda_k)(z^k - z^{k+1})$ (see (5.2.2)). The proof of this case is similar to the case $\beta f \mu_h > 0$ except that we use the above identity for $z^k$, the bound $\| (x_j^f - \gamma \nabla f(x_j^f)) - (x^* - \gamma \nabla f(x^*)) \|^2 \leq (1 + \gamma/\beta_f)^2 \| x_j^f - x^* \|^2$, and the constant $C'_2 = 3 \max \{ \gamma \lambda_k \mu_f, \gamma^2 / (\gamma \lambda_k (\beta_V - c \gamma)), (4/\lambda^2_k) \left( 2c - 1 \right) / \lambda^2_k \}$ in place of $C'_1$. Then the contraction $\| z^{k+1} - z^* \| \leq (1 - 1/C'_2)^{1/2} \| z^k - z^* \|$ follows.

In both cases, the linear rate for $(z^j)_{j \geq 0}$ follows by unfolding (5.6.1).

**Remark 5.6.1.** Note that smaller $c$ lead to larger $\gamma$ and smaller $(\lambda_j)_{j \geq 0}$, while larger $c$ lead to smaller $\gamma$ and larger $(\lambda_j)_{j \geq 0}$.

### 5.6.1 Arbitrarily slow convergence for strongly convex problems

In general, we cannot expect linear convergence of FDRS when $f$ is not differentiable—even if $f$ and $g$ are strongly convex. In this section, we construct an example to prove this claim. The following example is based on [6, Section 7] and Example 2.6.1 of Chapter 2.

**A family of slow examples**

Let $\mathcal{H} = \ell^2_2(N) = \mathbb{R}^2 \oplus \mathbb{R}^2 \oplus \cdots$. Let $R_{\theta}$ denote counterclockwise rotation in $\mathbb{R}^2$ by $\theta$ degrees. Let $e_0 := (1, 0)$ denote the standard unit vector, and let $e_{\theta} := R_{\theta} e_0$. Let $(\theta_j)_{j \geq 0}$ be a sequence of angles in $(0, \pi/2]$ such that $\theta_i \to 0$ as $i \to \infty$. For all $i \geq 0$, let $c_i := \cos(\theta_i)$. We let

$$V := \mathbb{R} e_0 \oplus \mathbb{R} e_0 \oplus \cdots \quad \text{and} \quad U := \mathbb{R} e_{\theta_i} \oplus \mathbb{R} e_{\theta_1} \oplus \cdots. \quad (5.6.3)$$

Note that [6, Section 7] proves the projection identities

$$(P_U)_i = \begin{bmatrix} \cos^2(\theta_i) & \sin(\theta_i) \cos(\theta_i) \\ \sin(\theta_i) \cos(\theta_i) & \sin^2(\theta_i) \end{bmatrix} \quad \text{and} \quad (P_V)_i = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

We now begin our extension of this example. Choose $a \geq 0$ and set $f = \chi_U + (a/2) \| \cdot \|^2$ and $g = (1/2) \| \cdot \|^2$. Note that $\mu_g = 1$ and $\mu_f = a$. In addition, for $h = g \circ P_V$, we have
\((\nabla h(x))_i = (P_V \circ I_H \circ P_V)_i = (P_V)_i\). Thus, \(\nabla h\) is 1-Lipschitz, and, hence, \(\beta_V = 1\) and we can choose \(\gamma = 1 < 2\beta_V\). Therefore, \(\alpha_{FDRS}^V = 2\beta_V/(4\beta_V - \gamma) = 2/3\), so we can choose \(\lambda_k \equiv 1 < 1/\alpha_{FDRS}^V\). We also note that \(\text{prox}_{\gamma f} = (1/(1 + a))P_U\).

Define \(N : \mathcal{H} \to \mathcal{H}\) on each 2-dimensional component of \(\mathcal{H}\) as follows: for all \(i \geq 0\),

\[
(N)_i := \left(\frac{1}{2}I_H + \frac{1}{2}\text{refl}_{\gamma f} \circ \text{refl}_{V}\right)_i = \frac{1}{a + 1}(P_U)_i(2(P_V)_i - I_{R^2}) + I_{R^2} - (P_V)_i
\]

\[
= \frac{1}{a + 1}(P_U)_i \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \frac{1}{a + 1} \begin{bmatrix} \cos^2(\theta_i) & -\sin(\theta_i)\cos(\theta_i) \\ \sin(\theta_i)\cos(\theta_i) & \cos^2(\theta_i) + a \end{bmatrix}
\]

where the second equality follows by direct expansion. Therefore, we have

\[
T_{FDRS} := N \circ (I - P_V) = \bigoplus_{i \geq 0} \frac{1}{a + 1} \begin{bmatrix} 0 & -\sin(\theta_i)\cos(\theta_i) \\ \cos^2(\theta_i) + a \end{bmatrix}.
\] (5.6.4)

Note that for all \(i \geq 0\), the operator \((T_{FDRS})_i\) has eigenvector

\[
\bar{z}_i = \left(\frac{-\cos(\theta_i)\sin(\theta_i)}{a + \cos^2(\theta_i)}, 1\right)
\]

with eigenvalue \(b_i := (a + c_i^2)/(a + 1) < 1\). Each component also has the eigenvector \((1, 0)\) with eigenvalue 0. Thus, the only fixed-point of \(T_{FDRS}\) is \(0 \in \mathcal{H}\). Finally,

\[
\|\bar{z}_i\|^2 = \frac{c_i^2(1 - c_i^2)}{(a + c_i^2)^2} + 1 \quad \text{and} \quad \|(P_V)_i\bar{z}_i\|^2 = \frac{c_i^2(1 - c_i^2)}{(a + c_i^2)^2}.
\] (5.6.5)

**Slow convergence proofs**

We know that \(z^{k+1} - z^k \to 0\) from (5.1.19). Therefore, because \(T_{FDRS}\) is linear, \([11,\ Proposition\ 5.27]\) proves the following lemma.

**Lemma 5.6.1** (Strong convergence for linear operators). Any sequence \((z^j)_{j \geq 0} \subseteq \mathcal{H}\) generated by the \(T_{FDRS}\) operator in (5.6.4) converges strongly to 0. Consequently, the sequences \((x^j_h)_{j \geq 0} = (P_V z^j)_{j \geq 0}\) and \((x^j_f)_{j \geq 0}\) converges strongly to zero.

**Lemma 5.6.2** (Slow sequences (Lemma 2.6.1 from Chapter 2)). Suppose that \(F : \mathbb{R}_+ \to (0, 1)\) is a function that is monotonically decreasing to zero. Then there exists a monotonic
sequence \((b_j)_{j \geq 0} \subseteq (0, 1)\) such that \(b_k \to 1^-\) as \(k \to \infty\) and an increasing sequence \((n_j)_{j \geq 0} \subseteq \mathbb{N} \cup \{0\}\) such that for all \(k \geq 0\),
\[
\frac{b_{n_k}^{k+1}}{(n_k + 1)} > e^{-1}F(k + 1).
\]

The following is a simple corollary of Lemma 5.6.2.

**Corollary 5.6.1.** Let the notation be as in Lemma 5.6.2. Then for all \(\eta \in (0, 1)\), we can find a sequence \((b_j)_{j \geq 0} \subseteq (\eta, 1)\) that satisfies the conditions of the lemma.

**Proof.** Replace the sequence \((b_j)_{j \geq 0}\) in Lemma 5.6.2 with \((\max\{b_j, \eta\})_{j \geq 0}\). \(\square\)

We are now ready to show that FDRS can converge arbitrarily slowly.

**Theorem 5.6.2** (Arbitrarily slow FDRS). For every function \(F : \mathbb{R}_+ \to (0, 1)\) that strictly decreases to zero, there is a point \(z^0 \in \ell_2^2(\mathbb{N})\) and two closed subspaces \(U\) and \(V\) with zero intersection, \(U \cap V = \{0\}\), such that the FDRS sequence \((z_j^i)_{j \geq 0}\) generated with the functions \(f = \chi_U + (a/2)\|\cdot\|^2\) and \(g = (1/2)\|\cdot\|^2\) and parameters \(\lambda_k \equiv 1\) and \(\gamma = 1\) strongly converges to zero, but for all \(k \geq 1\), we have
\[
\|z^k - z^*\| \geq e^{-1}F(k).
\]

**Proof.** For all \(i \geq 0\), define \(z^0_i = (1/\|z_i\|(i + 1))z_i\), then \(\|z^0_i\| = 1/(i + 1)\) and \(z^0_i\) is an eigenvector of \((T_{\text{FDRS}})_i\), with eigenvalue \(b_i = (a + c_i^2)/(a + 1)\). Define the concatenated vector \(z^0 = (z^0_i)_{i \geq 0}\). Note that \(z^0 \in \mathcal{H}\) because \(\|z^0\|^2 = \sum_{i=0}^{\infty} 1/(i + 1)^2 < \infty\). Thus, for all \(k \geq 0\), we let \(z^{k+1} = T_{\text{FDRS}}z^k\).

Now, recall that \(z^* = 0\). Thus, for all \(n \geq 0\) and \(k \geq 0\), we have
\[
\|z^{k+1} - z^*\|^2 = \|T_{\text{FDRS}}^n z^0\|^2 = \sum_{i=0}^{\infty} b_i^{2(k+1)}\|z^0_i\|^2 = \sum_{i=0}^{\infty} b_i^{2(k+1)}/(i + 1)^2 \geq \frac{b_n^{2(k+1)}}{(n + 1)^2}.
\]
Thus, \(\|z^{k+1} - z^*\| \geq b_n^{k+1}/(n+1)\). Choose \(b_n\) and the sequence \((n_j)_{j \geq 0}\) using Corollary 5.6.1 with \(\eta \in (a/(a + 1), 1)\). Then solve \(c_n = \sqrt{b_n(1 + a) - a} > 0\). \(\square\)
Remark 5.6.2. Theorems 5.6.2 and 5.4.1 show that the sequence $(z^j)_{j \geq 0}$ can converge arbitrarily slowly even if $(x^j_f)_{j \geq 0}$ and $(x^j_h)_{j \geq 0}$ converge with rate $o(1/\sqrt{k+1})$.

The following theorem shows that $(x^j_f)_{j \geq 0}$ and $(x^j_h)_{j \geq 0}$ do not converge linearly. See Appendix B.9 for the proof.

**Theorem 5.6.3.** There exists a sequence $(c_i)_{i \geq 0}$ so that $(x^j_h)_{j \geq 0}$ and $(x^j_f)_{j \geq 0}$ converge strongly, but not linearly. In particular, for any $\alpha > 1/2$, there is an initial point $z^0 \in H$ so that for all $k \geq 1$,

$$
\|x^k_h - x^*\|^2 \geq \frac{1}{(k+1)^{2\alpha}} \quad \text{and} \quad \|x^k_f - x^*\|^2 \geq \frac{(a+1/2)^2}{(a+1)^4(k+1)^{2\alpha}}.
$$

Thus, the nonergodic "best" convergence rates in Part 3 of Theorem 5.4.1 are sharp.

5.7 Primal-dual splittings

In this section, we reformulate FDRS as a primal-dual algorithm applied to the dual of the following problem: minimize$_{x \in V} f(x) + h(x)$.

**Lemma 5.7.1** (FDRS is a primal-dual algorithm). Let $\tau = 1/\gamma$, and suppose that $(z^j)_{j \geq 0}$ is generated by the FDRS algorithm with $\lambda_k \equiv 1$. For all $k \geq 0$, let $y^k := -\nabla \chi_V(x^k_h)$. Then for all $k \geq 0$, we have the recursive update rule:

$$
\begin{align*}
\begin{cases}
y^{k+1} = P_{V^\perp}(y^k - \tau x^k_f); \\
x^{k+1}_f = \text{prox}_{\gamma f}(x^k_f - \gamma \nabla h(x^k_f) + \gamma (2y^{k+1} - y^k)).
\end{cases}
\end{align*}
$$

**Proof.** Fix $k \geq 0$. By Lemma 5.2.1, $z^{k+1} = x^k_f - \gamma y^k$, so $(-1/\gamma)z^{k+1} = y^k - \tau x^k_f$. Thus, the formula for $(y^j)_{j \geq 0}$ follows from $y^{k+1} = -\nabla \chi_V(x^{k+1}_h) = -(1/\gamma)P_{V^\perp}z^{k+1}$.

Now observe that

$$
x^k_f = P_Vx^k_f + P_{V^\perp}x^k_f = P_V(z^{k+1} + \gamma y^k) + P_{V^\perp}(z^{k+1} + \gamma y^k) = x^{k+1}_h + \gamma(y^k - y^{k+1}).
$$
Furthermore, $\nabla h(x^f_k) = \nabla h(P_V x^f_k) = \nabla h(P_V(z^{k+1} + \gamma y^k)) = \nabla h(x^{k+1}_h)$. Thus,

$$
x^{k+1}_f \overset{(5.2.2)}{=} x^{k+1}_h - \gamma \left( \tilde{\nabla} \chi_V(x^{k+1}_h) + \nabla h(x^{k+1}_h) + \tilde{\nabla} f(x^{k+1}_f) \right)
= \prox_{\gamma f}(x^{k+1}_h - \gamma \nabla h(x^{k+1}_h) + \gamma y^{k+1})
= \prox_{\gamma f}(x^{k}_f - \gamma \nabla h(x^{k}_f) + \gamma(2y^{k+1} - y^k)).
$$

The algorithm in (5.7.1) is the primal-dual forward-backward algorithm of Vũ and Condat [112, 52] applied to the following dual problem: minimize $x \in V^\perp (f + h)^*(x)$ where $(f + h)^*(\cdot) = \sup_{x \in H} \langle x, \cdot \rangle - (f + h)(x)$ is the Legendre-Fenchel transform of $f + h$ [11, Definition 13.1]. For convergence, [112, Theorem 3.1] requires $\gamma \tau < 1$ and $2\beta_V > \left( \min\{1/\gamma, 1/\tau\} (1 - \sqrt{\gamma \tau}) \right)^{-1}$ whereas FDRS only requires $\gamma < 2\beta_V$.

Thus, the FDRS algorithm is a limiting case of Vũ and Condat’s algorithm, much like the DRS algorithm [85] is a limiting case of Chambolle and Pock’s primal-dual algorithm [36]. In addition, the convergence rate analysis in Section 5.3 cannot be subsumed by the convergence rate analysis of the primal-dual gap of Vũ and Condat’s algorithm in Chapter 4, which only applies when $\gamma \tau < 1$. The original FDRS paper did not show this connection [30, Remark 6.3 (iii)].

5.8 Conclusion

In this paper, we provided a comprehensive convergence rate analysis of the FDRS algorithm under general convexity, strong convexity, and Lipschitz differentiability assumptions. In almost all cases, the derived convergence rates are shown to be sharp. In addition, we showed that the FDRS algorithm is the limiting case of a recently developed primal-dual forward-backward operator splitting algorithm and, thus, clarify how it relates to existing algorithms. Future work on FDRS might evaluate the performance of the algorithm on realistic problems.
CHAPTER 6

A Three-Operator Splitting Scheme and its
Optimization Applications

6.1 Introduction

Operator splitting schemes reduce complex problems built from simple pieces into a series smaller subproblems which can be solved sequentially or in parallel. Since the 1950s they have been successfully applied to problems in PDE and control, but recent large-scale applications in machine learning, signal processing, and imaging have created a resurgence of interest in operator-splitting based algorithms. These algorithms often have very simple descriptions, are straightforward to implement on computers, and have (nearly) state-of-the-art performance for large-scale optimization problems. Although operator splitting techniques were introduced over 60 years ago, their importance has significantly increased in the past decade.

This chapter introduces a new operator-splitting scheme, which solves nonsmooth optimization problems of many different forms, as well as monotone inclusions. In an abstract form, this new splitting scheme will

\[
\text{find } x \in \mathcal{H} \text{ such that } 0 \in Ax + Bx + Cx
\]  

(6.1.1)

for three maximal monotone operators \( A, B, C \) defined on a Hilbert space \( \mathcal{H} \), where the \textbf{operator \( C \) is cocoercive}.\textsuperscript{1}

\textsuperscript{1}An operator \( C \) is \( \beta \)-cocoercive (or \( \beta \)-inverse-strongly monotone), \( \beta > 0 \), if \( \langle Cx - Cy, x - y \rangle \geq \beta \|Cx - Cy\|^2 \), \( \forall x, y \in \mathcal{H} \). This property generalizes many others. In particular, \( \nabla h \) of an \( L \)-Lipschitz differentiable convex function \( h \) is \( 1/L \)-cocoercive.
The most straightforward example of (6.1.1) arises from the optimization problem

\[
\text{minimize } f(x) + g(x) + h(x),
\]

where \( f, g, \) and \( h \) are proper, closed, and convex functions and \( h \) is \textbf{Lipschitz differentiable}. The first-order optimality condition of (6.1.2) reduces to (6.1.1) with \( Ax = \partial f(x), \) \( Bx = \partial g(x), \) and \( Cx = \nabla h(x), \) where \( \partial f, \partial g \) are subdifferentials of \( f \) and \( g, \) respectively. Note that \( C \) is cocoercive because \( h \) is Lipschitz differentiable.

A number of other examples of (6.1.1) can be found in Section 6.2 including split feasibility, doubly regularized, and monotropic programming problems, which have surprisingly many applications.

To introduce our splitting scheme, let \( I_H \) denote the identify map in \( H \) and \( J_S := (I + S)^{-1} \) denote the resolvent of a monotone operator \( S. \) (When \( S = \partial f, \) \( J_S(x) \) reduces to the proximal map: \( \arg \min_y f(y) + \frac{1}{2} ||x - y||^2. \) ) Let \( \gamma \in (0, 2\beta) \) be a scalar. Our splitting scheme for solving (6.1.1) is summarized by the operator

\[
T := I_H - J_{\gamma B} + J_{\gamma A} \circ (2J_{\gamma B} - I_H - \gamma C \circ J_{\gamma B}).
\]

Calculating \( Tx \) requires evaluating \( J_{\gamma A}, J_{\gamma B}, \) and \( C \) only once each, though \( J_{\gamma B} \) appears three times in \( T. \) In addition, we will show that a fixed-point of \( T \) encodes a solution to (6.1.1) and \( T \) is an averaged operator.

The problem (6.1.1) can be solved by iterating

\[
z^{k+1} := (1 - \lambda_k)z^k + \lambda_k Tz^k,
\]

where \( z^0 \) is an arbitrary point and \( \lambda_k \in (0, (4\beta - \gamma)/2\beta) \) is a relaxation parameter. (For simplicity, one can fix \( \gamma < 2\beta \) and \( \lambda_k \equiv 1. \) ) This iteration can be implemented as follows:

\textbf{Algorithm 6.1.1.} Set an arbitrary point \( z^0 \in \mathcal{H}, \) stepsize \( \gamma \in (0, 2\beta), \) and relaxation sequence \( (\lambda_j)_{j \geq 0} \in (0, (4\beta - \gamma)/2\beta). \) For \( k = 0, 1, \ldots, \) iterate:

1. get \( x^k_B = J_{\gamma B}(z^k); \)
2. get

\[ x^k_A = J_{\gamma A}(2x^k_B - z^k - \gamma Cx^k_B); \quad \text{//comment: } x^k_A = J_{\gamma A} \circ (2J_{\gamma B} - I_H - \gamma C \circ J_{\gamma B})z^k \]

3. get

\[ z^{k+1} = z^k + \lambda_k (x^k_A - x^k_B); \quad \text{//comment: } z^{k+1} = (1 - \lambda_k)z^k + \lambda_k Tz^k \]

Algorithm 6.1.1 leads to new algorithms for a large number of applications, which are given in Section 6.2 below. Although some of those applications can be solved by other splitting methods, for example, by the alternating directions method of multipliers (ADMM), our new algorithms are typically simpler, use fewer or no additional variables, and take advantage of the differentiability of smooth terms in the objective function. The dual form of our algorithm is the simplest extension of ADMM from the classic two-block form to the three-block form that has a general convergence result. The details of these are given in Section 6.2.

The full convergence result for Algorithm 6.1.1 is stated in Theorem 6.3.1. For brevity we include the following simpler version here:

**Theorem 6.1.1** (Convergence of Algorithm 6.1.1). Suppose that \( \text{Fix } T \neq \emptyset \). Let \( \alpha = 2\beta/(4\beta - \gamma) \) and suppose that \( (\lambda_j)_{j \geq 0} \) satisfies \( \sum_{j=0}^{\infty} (1 - \lambda_j/\alpha)\lambda_j/\alpha = \infty \) (which is true if the sequence is strictly bounded away from 0 and 1/\( \alpha \)). Then the sequences \( (z^j)_{j \geq 0} \), \( (x^j_B)_{j \geq 0} \), and \( (x^j_A)_{j \geq 0} \) generated by Algorithm 6.1.1 satisfy the following:

1. \( (z^j)_{j \geq 0} \) converges weakly to a fixed point of \( T \); and
2. \( (x^j_B)_{j \geq 0} \) and \( (x^j_A)_{j \geq 0} \) converge weakly to an element of \( \text{zer}(A + B + C) \).

### 6.1.1 Existing two–operator splitting schemes

A large variety of recent algorithms [36, 61, 104] and their generalizations and enhancements [20, 24, 23, 29, 44, 45, 47, 52, 78, 112] are (skillful) applications of one of the following three operator-splitting schemes: (i) forward-backward-forward splitting (FBFS) [111], (ii) forward-backward splitting (FBS) [100], and (iii) Douglas-Rachford splitting (DRS) [85], which all split the sum of two operators. (The recently introduced forward-Douglas-Rachford splitting (FDRS) turns out to be a special case of FBS applied to a suitable
monotone inclusion as shown in Section 5.7 of Chapter 5.) Until now, these algorithms are
the only basic operator-splitting schemes for monotone inclusions, if we ignore variants
involving inertial dynamics, special metrics, Bregman divergences, or different stepsize
choices\(^2\). To our knowledge, no new splitting schemes have been proposed since the in-
troduction of FBFS in 2000.

The proposed splitting scheme \( T \) in Equation (6.1.3) is the first algorithm to split the
sum of three operators that does not appear to reduce to any of the existing schemes. In
fact, FBS, DRS, and FDRS are special cases of Algorithm 6.1.1.

**Proposition 6.1.1** (Existing operator splitting schemes as special cases).  1. Consider
the forward-backward splitting (FBS) operator [100], \( T_{\text{FBS}} := J_{\gamma A} \circ (I_H - \gamma C) \), for
solving \( 0 \in Ax + Cx \) where \( A \) is maximal monotone and \( C \) is cocoercive. If we set
\( B = 0 \) in (6.1.3), then \( T = T_{\text{FBS}} \).

2. Consider the Douglas-Rachford splitting (DRS) operator [85], \( T_{\text{DRS}} := I_H - J_{\gamma B} + J_{\gamma A} \circ (2J_{\gamma B} - I_H) \), for solving \( 0 \in Ax + Bx \) where \( A, B \) are maximal monotone. If
we set \( C = 0 \) in (6.1.3), then \( T = T_{\text{DRS}} \).

3. Consider the forward-Douglas-Rachford splitting (FDRS) operator [30], \( T_{\text{FDRS}} := I_H - P_V + J_{\gamma A} \circ (2P_V - I_H - \gamma P_V \circ C' \circ P_V) \), for solving \( 0 \in Ax + C'x + N_V x \) where \( A \) is maximal monotone, \( C' \) is cocoercive, \( V \) is a closed vector space, \( N_V \) is the normal
cone operator of \( V \), and \( P_V \) denote the projection to \( V \). If we set \( B = N_V \) and
\( C = P_V \circ C' \circ P_V \) in (6.1.3), then \( T = T_{\text{FDRS}} \).

The operator \( T \) is also related to the Peaceman-Rachford splitting (PRS) operator
[85]. Let us introduce the “reflection” operator \( \text{refl}_A := 2J_A - I_H \) where \( A : \mathcal{H} \to \mathcal{H} \) is a
maximal monotone operator, and set

\[
S := 2T - I = \text{refl}_{\gamma A} \circ (\text{refl}_{\gamma B} - \gamma C \circ J_{\gamma B}) - \gamma C \circ J_{\gamma B}.
\]

\(^2\)For example, Peaceman-Rachford splitting (PRS) [85] doubles the step size in DRS.
If we set $C = 0$, then $S$ reduces to the PRS operator.

### 6.1.2 Convergence rate guarantees

We show in Lemma 6.3.2 that from any fixed point $z^*$ of the operator $T$, we obtain $x^* := J_{\gamma B}(z^*)$ as a zero of the monotone inclusion (6.1.1), i.e., $x^* \in \text{zer}(A + B + C)$. In addition, under various scenarios, the following convergence rates can be deduced:

1. **Fixed-point residual (FPR) rate**: The FPR $\|T z^k - z^k\|^2$ has the sharp rate $o \left( \frac{1}{\sqrt{k + 1}} \right)$. (Part 7 of Theorem 6.3.1 and Remark 6.3.5.)

2. **Function value rate**: Under mild conditions on Problem (6.1.2), although $(f + g + h)(x^k) - (f + g + h)(x^*)$ is not monotonic, it is bounded by $o \left( \frac{1}{\sqrt{k + 1}} \right)$. Two averaging procedures improve this rate to $O \left( \frac{1}{(k + 1)} \right)$. The running best sequence, $\min_{i=0, \ldots, k} (f + g + h)(x^i) - (f + g + h)(x^*)$, further improves to $o \left( \frac{1}{(k + 1)} \right)$ whenever $f$ is differentiable and $\nabla f$ is Lipschitz continuous. These rates are also sharp.

3. **Strong convergence**: When $A$ (respectively $B$ or $C$) is strongly monotone, the sequence $\|x^A_A - x^*\|^2$ (respectively $\|x^B_B - x^*\|^2$) converges with rate $o(1/\sqrt{k + 1})$. The running best and averaged sequences improve this rate to $o(1/(k + 1))$ and $O(1/(k + 1))$, respectively.

4. **Linear convergence**: We reserve $\mu \in [0, \infty)$ for strong monotonicity constants and $L \in (0, \infty]$ for Lipschitz constants. If strong monotonicity does not hold, then $\mu = 0$. If Lipschitz continuity does not hold, then $L = \infty$. Algorithm 6.1.1 converges linearly whenever $(\mu_A + \mu_B + \mu_C)(1/L_A + 1/L_B) > 0$, i.e., whenever at least one of $A$, $B$, or $C$ is strongly monotone and at least one of $A$ or $B$ is Lipschitz continuous. We present a counterexample where $A$ and $B$ are not Lipschitz continuous and Algorithm 6.1.1 fails to converge linearly.

5. **Variational inequality convergence rate**: We can apply Algorithm 6.1.1 to primal-dual optimality conditions and other structured monotone inclusions with...
\( A = \mathcal{A} + \partial f \), \( B = \mathcal{B} + \partial g \) and \( C = \mathcal{C} + \nabla h \) for some monotone operators \( \mathcal{A} \), \( \mathcal{B} \), and \( \mathcal{C} \). A typical example is when \( \mathcal{A} \) and \( \mathcal{B} \) are bounded skew linear maps and \( \mathcal{C} = 0 \). Then, the corresponding variational inequality converges with rate \( o \left( \frac{1}{\sqrt{k+1}} \right) \) under mild conditions on \( \mathcal{A} \) and \( f \). Again, averaging can improve the rate to \( O(1/(k+1)) \).

### 6.1.3 Modifications and enhancements of the algorithm

#### 6.1.3.1 Averaging

The averaging strategies in this subsection maintain additional running averages of its sequences \((x^j_A)_{j \geq 0}\) and \((x^j_B)_{j \geq 0}\) in Algorithm 6.1.1. Compared to the worst-case rate \( o(1/\sqrt{k+1}) \) of the original iterates, the running averages have the improved rate of \( O(1/(k+1)) \), which is referred to as the *ergodic rate*. This better rate, however, is often contradicted by worse practical performance, for the following reasons: (i) In many finite dimensional applications, when the iterates reach a solution neighborhood, convergence improves from sublinear to linear, but the ergodic rate typically stays sublinear at \( O(1/(k+1)) \); (ii) structures such as sparsity and low-rankness in current iterates often get lost when they are averaged with all their past iterates. This effect is dramatic in sparse optimization because the average of many sparse vectors can be dense.

The following averaging scheme is typically used in the literature for splitting schemes (See Chapters 2, 3, 4, and 5 and reference [20]):

\[
\bar{x}_B^k = \frac{1}{\sum_{i=0}^{k} \lambda_i} \sum_{i=0}^{k} \lambda_i x_i^B \quad \text{and} \quad \bar{x}_A^k = \frac{1}{\sum_{i=0}^{k} \lambda_i} \sum_{i=0}^{k} \lambda_i x_i^A, \tag{6.1.6}
\]

where all \( \lambda_i, x_i^A, \) and \( x_i^B \) are given by Algorithm 6.1.1. By maintaining the running averages in Algorithm 6.1.1, \( \bar{x}_B^k \) and \( \bar{x}_A^k \) are essentially costless to compute.

The following averaging scheme, inspired by [92], uses a *constant sequence of relaxation*...
parameters $\lambda_i$ but it gives more weight to the later iterates:

$$
\overline{x}_B^k = \frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} (i+1)x_B^i \quad \text{and} \quad \overline{x}_A^k = \frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} (i+1)x_A^i. \quad (6.1.7)
$$

This seems intuitively better: the older iterates should matter less than the current iterates. The above ergodic iterates are closer to the current iterate, but they maintain the improved convergence rate of $O(1/(k+1))$. Like before, $\overline{x}_B^k$ and $\overline{x}_A^k$ can be computed by updating $\overline{x}_B^{k-1}$ and $\overline{x}_A^{k-1}$ at little cost.

### 6.1.3.2 Some accelerations

In this section we introduce an acceleration of Algorithm 6.1.1 that applies whenever $B$ or $C$ is strongly monotone. If $f$ is strongly convex, then $S = \partial f$ is strongly monotone. Instead of fixing the step size $\gamma$, a *varying* sequence of stepsizes $(\gamma_j)_{j \geq 0}$ are used for acceleration. The acceleration is significant on problems where Algorithm 6.1.1 works nearly at its performance lower bound and the strong convexity constants are easy to obtain. The new algorithm is presented in variables different from those in Algorithm 6.1.1 since the change from $\gamma_k$ to $\gamma_{k+1}$ occurs in the middle of each iteration of Algorithm 6.1.1, right after $J_{\gamma_B}$ is applied. In case that $\gamma_k \equiv \gamma$ is fixed, the new algorithm reduces to Algorithm 6.1.1 with a constant relaxation parameter $\lambda_k \equiv 1$ via the change of variable: $z^k = x_A^{k-1} + \gamma_k u_B^{k-1}$. The new algorithm is as follows:

**Algorithm 6.1.2** (Algorithm 6.1.1 with acceleration). *Choose* $z^0 \in \mathcal{H}$ *and stepsizes* $(\gamma_j)_{j \geq 0} \in (0, \infty)$. *Let* $x_A^0 \in \mathcal{H}$ *and set* $x_B^0 = J_{\gamma_B}(x_A^0), u_B^0 = (1/\gamma_0)(I - J_{\gamma_B})(x_A^0)$. *For* $k = 1, 2, \ldots$, *iterate*

1. get $x_B^k = J_{\gamma_B}(x_A^{k-1} + \gamma_k u_B^{k-1})$;
2. get $u_B^k = (1/\gamma_{k-1})(x_A^{k-1} + \gamma_k u_B^{k-1} - x_B^k)$;
3. get $x_A^k = J_{\gamma_k A}(x_B^k - \gamma_k u_B^k - \gamma_k C x_B^k)$;
The sequence of stepsizes \((\gamma_j)_{j \geq 0}\), which are related to [36, Algorithm 2] and [21, Algorithm 5], are introduced in Theorem 6.1.2. These stepsizes improve the convergence rate of \(\|x^k_B - x^*\|^2\) to \(O(1/(k+1)^2)\).

**Theorem 6.1.2** (Accelerated variants of Algorithm 6.1.1). Let \(B\) be \(\mu_B\)-strongly monotone, where we allow the case \(\mu_B = 0\).

1. Suppose that \(C\) is \(\beta\)-cocoercive and \(\mu_C\)-strongly monotone. Let \(\eta \in (0, 1)\) and choose \(\gamma_0 \in (0, 2\beta(1 - \eta))\). In algorithm 6.1.2, for all \(k \geq 0\), let
   \[
   \gamma_{k+1} := \frac{-2\gamma_k^2 \mu_C \eta + \sqrt{(2\gamma_k^2 \mu_C \eta)^2 + 4(1 + 2\gamma_k \mu_B)\gamma_k^2}}{2(1 + 2\gamma_k \mu_B)}.
   \]
   Then we have \(\|x^k_B - x^*\|^2 = O(1/(k+1)^2)\).

2. Suppose that \(C\) is \(L_C\)-Lipschitz, but not necessarily strongly monotone or cocoercive. Suppose that \(\mu_B > 0\). Let \(\gamma_0 \in (0, 2\mu_B/L_C^2)\). In algorithm 6.1.2, for all \(k \geq 0\), let
   \[
   \gamma_{k+1} := \frac{\gamma_k}{\sqrt{1 + 2\gamma_k (\mu_B - \gamma_k L_C^2/2)}}
   \]
   Then we have \(\|x^k_B - x^*\|^2 = O(1/(k+1)^2)\).

The proof can be found in Appendix C.1.

### 6.1.4 Practical implementation issues: Line search

Recall that \(\beta\), the cocoercivity constant of \(C\), determines the stepsize condition \(\gamma \in (0, 2\beta)\) for Algorithm 6.1.1. When \(\beta\) is unknown, one can find \(\gamma\) by trial and error. Whenever the FPR is observed to increase (which does not happen if \(\gamma \in (0, 2\beta)\) by Part 2 of Theorem 6.3.1), reduce \(\gamma\) and restart the algorithm from the initial or last iterate.

For the case of \(C = \nabla h\) for some convex function \(h\) with Lipschitz \(\nabla h\), we propose a line search procedure that uses a fixed stepsize \(\gamma\) but involves an auxiliary factor \(\rho \in (0, 1]\). It works better than the above approach of changing \(\gamma\) since the latter changes fixed point. Let

\[
\text{refl}^\rho_{\gamma_B} := (1 + \rho)J_{\gamma_B} - \rho I_H.
\]
Note that $\text{refl}_{\gamma_B}^1 = \text{refl}_{\gamma_B}$ and $\text{refl}_{\gamma_B}^0 = J_{\gamma_B}$. Define

$$T_{\gamma}^\rho = I_H - J_{\gamma_B} + J_{\rho \gamma_A} \circ (\text{refl}_{\gamma_B}^\rho - \rho \gamma \nabla h \circ J_{\gamma_B}).$$

Our line search procedure iterates $z^{k+1} = T_{\gamma}^\rho(z^k)$ with a special choice of $\rho$:

**Algorithm 6.1.3** (Algorithm 6.1.1 with line search). Choose $z^0 \in H$ and $\gamma \in (0, \infty)$.

For $k = 0, 1, \ldots$, iterate

1. get $x_B^k = J_{\gamma_B}(z^k)$;

2. get $\rho \in (0, 1]$ such that

$$h(x_A^k) \leq h(x_B^k) + \langle x_A^k - x_B^k, \nabla h(x_B^k) \rangle + \frac{1}{2\gamma \rho} \|x_A^k - x_B^k\|^2$$

where

$$x_A^k = J_{\rho \gamma_A}(x_B^k + \rho(x_B^k - z^k) - \gamma \rho \nabla h(x_B^k));$$

3. get $z^{k+1} = z^k + x_A^k - x_B^k$.

A straightforward calculation shows the following lemma:

**Lemma 6.1.1.** For all $\rho \in (0, 1)$ and all $\gamma > 0$, we have

$$\text{zer}(A + B + \nabla h) = J_{\gamma_B}(\text{Fix } T_{\gamma}^\rho) \quad \text{and} \quad \text{Fix}(T_{\gamma}^\rho) = \text{Fix}(T_{\gamma}^1).$$

**Remark 6.1.1.** In practice, Algorithm 6.1.3, which can start with a larger $\gamma$, can be an order of magnitude faster than Algorithm 6.1.1. Unfortunately, we have no proof of convergence for this method.

### 6.1.5 Definitions, notation and some facts

In what follows, $H$ denotes a (possibly infinite dimensional) Hilbert space. We use $\langle , \rangle$ to denote the inner product associated to a Hilbert space. In all of the algorithms we consider, we utilize two stepsize sequences: the implicit sequence $(\gamma_j)_{j \geq 0} \subseteq \mathbb{R}_{++}$ and the explicit sequence $(\lambda_j)_{j \geq 0} \subseteq \mathbb{R}_{++}$. 

218
The following definitions and facts are mostly standard and can be found in [11].

Let \( L \geq 0 \), and let \( D \) be a nonempty subset of \( \mathcal{H} \). A map \( T : D \to \mathcal{H} \) is called \( L \)-Lipschitz if for all \( x, y \in \mathcal{H} \), we have \( \|Tx - Ty\| \leq L\|x - y\| \). In particular, \( N \) is called nonexpansive if it is 1-Lipschitz. A map \( N : D \to \mathcal{H} \) is called \( \lambda \)-averaged [11, Section 4.4] if it can be written as

\[
N = T_\lambda := (1 - \lambda)I_{\mathcal{H}} + \lambda T
\]  

(6.1.10)

for a nonexpansive map \( T : D \to \mathcal{H} \) and a real number \( \lambda \in (0, 1) \). A \((1/2)\)-averaged map is called firmly nonexpansive. We use a * superscript to denote a fixed point of a nonexpansive map, e.g., \( z^* \).

Let \( 2^\mathcal{H} \) denote the power set of \( \mathcal{H} \). A set-valued operator \( A : \mathcal{H} \to 2^\mathcal{H} \) is called monotone if for all \( x, y \in \mathcal{H} \), \( u \in Ax \), and \( v \in Ay \), we have \( \langle x - y, u - v \rangle \geq 0 \). We denote the set of zeros of a monotone operator by \( \text{zer}(A) := \{ x \in \mathcal{H} \mid 0 \in Ax \} \). The graph of \( A \) is denoted by \( \text{gra}(A) := \{(x, y) \mid x \in \mathcal{H}, y \in Ax\} \). Evidently, \( A \) is uniquely determined by its graph. A monotone operator \( A \) is called maximal monotone provided that \( \text{gra}(A) \) is not properly contained in the graph of any other monotone set-valued operator. The inverse of \( A \), denoted by \( A^{-1} \), is defined uniquely by its graph \( \text{gra}(A^{-1}) := \{(y, x) \mid x \in \mathcal{H}, y \in Ax\} \).

Let \( \beta \in \mathbb{R} \) be a positive real number. The operator \( A \) is called \( \beta \)-strongly monotone provided that for all \( x, y \in \mathcal{H} \), \( u \in Ax \), and \( v \in Ay \), we have \( \langle x - y, u - v \rangle \geq \beta \|x - y\|^2 \).

A single-valued operator \( B : \mathcal{H} \to 2^\mathcal{H} \) maps each point in \( \mathcal{H} \) to a singleton and will be identified with the natural \( \mathcal{H} \)-valued map it defines. The resolvent of a monotone operator \( A \) is defined by the inversion \( J_A := (I + A)^{-1} \). Minty’s theorem shows that \( J_A \) is single-valued and has full domain \( \mathcal{H} \) if, and only if, \( A \) is maximally monotone. Note that \( A \) is monotone if, and only if, \( J_A \) is firmly nonexpansive. Thus, the reflection operator

\[
\text{refl}_A := 2J_A - I_{\mathcal{H}}
\]  

(6.1.11)

is nonexpansive on \( \mathcal{H} \) whenever \( A \) is maximally monotone.

Let \( f : \mathcal{H} \to (-\infty, \infty] \) denote a closed (i.e., lower semi-continuous), proper, and convex function. Let \( \text{dom}(f) := \{ x \in \mathcal{H} \mid f(x) < \infty \} \). We let \( \partial f(x) : \mathcal{H} \to 2^\mathcal{H} \) denote the
subdifferential of \( f \): \( \partial f(x) := \{ u \in H \mid \forall y \in H, f(y) \geq f(x) + \langle y - x, u \rangle \} \). We always let
\[
\tilde{\nabla} f(x) \in \partial f(x)
\]
denote a subgradient of \( f \) drawn at the point \( x \). The subdifferential operator of \( f \) is maximally monotone. The inverse of \( \partial f \) is given by \( \partial f^* \) where \( f^*(y) := \sup_{x \in H} \langle y, x \rangle - f(x) \) is the Fenchel conjugate of \( f \). If the function \( f \) is \( \beta \)-strongly convex, then \( \partial f \) is \( \beta \)-strongly monotone and \( \partial f^* \) is single-valued and \( \beta \)-cocoercive.

If a convex function \( f : H \to (-\infty, \infty] \) is Fréchet differentiable at \( x \in H \), then \( \partial f(x) = \{ \nabla f(x) \} \). Suppose \( f \) is convex and Fréchet differentiable on \( H \), and let \( \beta \in \mathbb{R} \) be a positive real number. Then the Baillon-Haddad theorem states that \( \nabla f \) is \( (1/\beta) \)-Lipschitz if, and only if, \( \nabla f \) is \( \beta \)-cocoercive.

The resolvent operator associated to \( \partial f \) is called the proximal operator and is uniquely defined by the following (strongly convex) minimization problem: \( \text{prox}_f(x) := J_{\partial f}(x) = \arg \min_{y \in H} f(y) + (1/2)\|y - x\|^2 \). The indicator function of a closed, convex set \( C \subseteq H \) is denoted by \( \iota_C : H \to \{0, \infty\} \); the indicator function is 0 on \( C \) and is \( \infty \) on \( H \setminus C \). The normal cone operator of \( C \) is the monotone operator \( N_C := \partial \iota_C \).

Finally, we call the following identity the cosine rule:
\[
\|y - z\|^2 + 2\langle y - x, z - x \rangle = \|y - x\|^2 + \|z - x\|^2, \quad \forall x, y, z \in H. \tag{6.1.12}
\]

### 6.2 Motivation and Applications

Our splitting scheme provides simple numerical solutions to a large number of problems that appear in signal processing, machine learning, and statistics. In this section, we provide some concrete problems that reduce to the monotone inclusion problem (6.1.1). These are a small fraction of the problems to which our algorithm will apply. For example, when a problem has four or more blocks, we can reduce it to three or fewer blocks by grouping similar components or lifting the problem to a higher-dimensional space.

For every method, we list the three monotone operators \( A, B, \) and \( C \) from prob-
lem (6.1.1), and a minimal list of conditions needed to guarantee convergence.

We do not include any examples with only one or two blocks they can be solved by existing splitting algorithms that are special cases of our algorithm.

6.2.1 The 3-set (split) feasibility problem

This problem is to find

\[ x \in C_1 \cap C_2 \cap C_3, \tag{6.2.1} \]

where \( C_1, C_2, C_3 \) are three nonempty convex sets and the projection to each set can be computed numerically. The more general 3-set split feasibility problem is to find

\[ x \in C_1 \cap C_2 \text{ such that } \ Lx \in C_3, \tag{6.2.2} \]

where \( L \) is a linear mapping. We can reformulate the problem as

\[
\min_{x} \frac{1}{2} d^2(Lx, C_3) \text{ subject to } x \in C_1 \cap C_2, \tag{6.2.3}
\]

where \( d(Lx, C_3) := \|Lx - P_{C_3}(Lx)\| \) and \( P_{C_3} \) denotes the projection to \( C_3 \). Problem (6.2.2) has a solution if and only if problem (6.2.3) has a solution that gives 0 objective value.

The following algorithm is an instance of Algorithm 6.1.1 applied with the monotone operators:

\[
A x := N_{C_1}(x); \quad B x := N_{C_2}(x); \quad C x := \nabla_{x} \frac{1}{2} d^2(Lx, C_3) = L^*(Lx - P_{C_3}(Lx)).
\]

**Algorithm 6.2.1** (3-set split feasibility algorithm). *Set an arbitrary \( z^0 \in \mathcal{H} \), stepsize \( \gamma \in (0, 2/\|L\|^2) \), and sequence of relaxation parameters \( (\lambda_j)_{j \geq 0} \in (0, 2 - \gamma\|L\|^2/2) \). For \( k = 0, 1, \ldots \), iterate

1. get \( x^k = P_{C_2}(z^k) \);
2. get \( y^k = Lx^k \);
3. get \( z^{k+\frac{1}{2}} = 2x^k - z^k - \gamma L^*(y^k - P_{C_3}(y^k)) \); \quad //comment: \( z^{k+\frac{1}{2}} = (2J_{\gamma B} - I_{\mathcal{H}} - \gamma C \circ J_{\gamma_B})z^k \)
4. get $z^{k+1} = z^k + \lambda_k (P_{C_1}(z^{k+\frac{1}{2}}) - x^k)$.

Note that the algorithm only explicitly applies $L$ and $L^*$, the adjoint of $L$, and does not need to invert a map involving $L$ or $L^*$. The stepsize rule $\gamma \in (0, 2/\|L\|^2)$ follows because $\nabla_{x^\frac{1}{2}}d^2(x, C_3)$ is 1-Lipschitz [11, Corollary 12.30].

6.2.2 The 3-objective minimization problem

The problem is to find a solution to

$$\min_{x} f(x) + g(x) + h(Lx),$$

(6.2.4)

where $f, g, h$ are proper closed convex functions, $h$ is $(1/\beta)$-Lipschitz-differentiable, and $L$ is a linear mapping. Note that any constraint $x \in C$ can be written as the indicator function $\iota_C(x)$ and incorporated in $f$ or $g$. Therefore, the problem (6.2.3) is a special case of (6.2.4).

The following algorithm is an instance of Algorithm 6.1.1 applied with the monotone operators:

$$A = \partial f; \quad B = \partial g; \quad C = \nabla(h \circ L) = L^* \circ \nabla h \circ L.$$

Algorithm 6.2.2 (for problem (6.2.4)). Set an arbitrary $z^0$, stepsize $\gamma \in (0, 2/(\beta \|L\|^2))$, and sequence of relaxation parameters $(\lambda_j)_{j \geq 0} \in (0, 2 - \gamma \beta \|L\|^2/2)$. For $k = 0, 1, \ldots,$ iterate

1. get $x^k = \prox_{\gamma g}(z^k)$;

2. get $y^k = Lx^k$;

3. get $z^{k+\frac{1}{2}} = 2x^k - z^k - \gamma L^* \nabla h(y^k)$; //comment: $z^{k+\frac{1}{2}} = (2J_B - I_H - \gamma C \circ J_B)z^k$

4. get $z^{k+1} = z^k + \lambda_k (\prox_{\gamma f}(z^{k+\frac{1}{2}}) - x^k)$. 

222
6.2.2.1 Application: double-regularization and multi-regularization

Regularization helps recover a signal with the structures that are either known a priori or sought after. In practice, regularization is often enforced through nonsmooth objective functions, such as $\ell_1$ and nuclear norms, or constraints, such as nonnegativity, bound, linear, and norm constraints. Many problems involve more than one regularization term (counting both objective functions and constraints), in order to reduce the “search space” and more accurately shape their solutions. Such problems have the general form

$$\min_{x \in H} \sum_{i=1}^{m} r_i(x) + h_0(Lx),$$

(6.2.5)

where $r_i$ are possibly-nonsmooth regularization functions and $h_0$ is a Lipschitz differentiable function. When $m = 1, 2$, our algorithms can be directly applied to (6.2.5) by setting $f = r_1$ and $g = r_2$ in Algorithm 6.2.2.

When $m \geq 3$, a simple approach is to introduce variables $x^{(i)}$, $i = 1, \ldots, m$, and apply Algorithm 6.2.2 to either of the following problems, both of which are equivalent to (6.2.5):

$$\min_{x, x^{(1)}, \ldots, x^{(m)} \in H} \sum_{i=1}^{m} r_i(x^{(i)}) + \underbrace{t_{\{x=x^{(1)}=\cdots=x^{(m)}\}}(x, x^{(1)}, \ldots, x^{(m)}) + h_0(Lx)}_{g},$$

(6.2.6)

$$\min_{x^{(1)}, \ldots, x^{(m)} \in H} \sum_{i=1}^{m} \left( r_i(x^{(i)}) + \frac{1}{m} h_0(Lx^{(i)}) \right) + \underbrace{t_{\{x=x^{(1)}=\cdots=x^{(m)}\}}(x^{(1)}, \ldots, x^{(m)})}_{g},$$

(6.2.7)

where $g$ returns 0 if all the inputs are identical and $\infty$ otherwise. Problem (6.2.6) has a simpler form, but problem (6.2.7) requires fewer variables and will be strongly convex in the product space whenever $h_0(Lx)$ is strongly convex in $x$.

It is easy to adapt Algorithm 6.2.2 for problems (6.2.7) and (6.2.6). We give the one for problem (6.2.7):

**Algorithm 6.2.3** (for problem (6.2.7)). Choose points $z_0^{(1)}, \ldots, z_0^{(m)}$, implicit stepsize $\gamma \in (0, 2m/(\beta\|L\|^2))$, and relaxation parameters $(\lambda_j)_{j \geq 0} \in (0, 2 - \gamma \beta \|L\|^2/(2m))$. For $k = 0, 1, \ldots$, iterate
1. get $x_k^{(1)}, \ldots, x_k^{(m)} = \frac{1}{m} (z_k^{(1)} + \cdots + z_k^{(m)});$ 

2. get $z_{(i)}^{k+1/2} = 2x_k^{(i)} - z_{(i)}^{k} - \frac{1}{m} L^* \nabla h(Lx_k^{(i)})$ for $i = 1, \ldots, m$, in parallel. 

3. get $z_{(i)}^{k+1} = z_{(i)}^{k} + \lambda_k \left( \text{prox}_{\gamma r} (z_{(i)}^{k+1/2} - x_k^{(i)}) \right)$ for $i = 1, \ldots, m$, in parallel. 

Because Step 1 yields identical $x_k^{(1)}, \ldots, x_k^{(m)}$, they can be consolidated to a single $x_k$ in both steps. For the same reason, splitting $h_0(L \cdot)$ into multiple copies does not incur more computation.

### 6.2.2.2 Application: texture inpainting

Let $y$ be a color texture image represented as a 3-way tensor where $y(:,:,1), y(:,:,2), y(:,:,3)$ are the red, green, and blue channels of the image, respectively. Let $P_\Omega$ be the linear operator that selects the set of known entries of $y$, that is, $P_\Omega y$ is given. The inpainting problem is to recover a set of unknown entries of $y$. Because the matrix unfoldings of the texture image $y$ are (nearly) low-rank (as in [86, Equation (4)]), we formulate the inpainting problem as

$$
\text{minimize } \omega \| x_{(1)} \|_* + \omega \| x_{(2)} \|_* + \frac{1}{2} \| P_\Omega x - P_\Omega y \|^2
$$

where $x$ is the 3-way tensor variable, $x_{(1)}$ is the matrix $[x(:,:,1) x(:,:,2) x(:,:,3)]$, $x_{(2)}$ is the matrix $[x(:,:,1)^T x(:,:,2)^T x(:,:,3)^T]^T$, $\| \cdot \|_*$ denotes matrix nuclear norm, and $\omega$ is a penalty parameter. Problem (6.2.8) can be solved by Algorithm 6.2.2. The proximal mapping of the term $\| \cdot \|_*$ can be computed by singular value soft-thresholding. Our numerical results are given in Section 6.5.1.

### 6.2.2.3 Matrix completion

Let $X_0 \in \mathbb{R}^{m \times n}$ be a matrix with entries that lie in the interval $[l, u]$, where $l < u$ are positive real numbers. Let $A$ be a linear map that “selects” a subset of the entries of an $m \times n$ matrix by setting each unknown entry in the matrix to 0. We are interested in
recovering matrices $X_0$ from the matrix of “known” entries $A(X_0)$. Mathematically, one approach to solve this problem is as follows [34]:

$$\minimize_{X \in \mathbb{R}^{m \times n}} \frac{1}{2} \| A(X - X_0) \|^2 + \mu \| X \|_*$$

subject to: $l \leq X \leq u$ \hspace{1cm} (6.2.9)

where $\mu > 0$ is a parameter, $\| \cdot \|$ is the Frobenius norm, and $\| \cdot \|_*$ is the nuclear norm. Problem (6.2.9) can be solved by Algorithm 6.2.2. The proximal operator of $\| \cdot \|_*$ ball can be computed by soft thresholding the singular values of $X$. Our numerical results are given in Section 6.5.2.

6.2.2.4 Application: support vector machine classification and portfolio optimization

Consider the constrained quadratic program in $\mathbb{R}^d$:

$$\minimize_{x \in \mathbb{R}^d} \frac{1}{2} \langle Qx, x \rangle + \langle c, x \rangle$$ \hspace{1cm} (6.2.10)

subject to $x \in C_1 \cap C_2$

where $Q \in \mathbb{R}^{d \times d}$ is a symmetric positive semi-definite matrix, $c \in \mathbb{R}^d$ is a vector, and $C_1, C_2 \subseteq \mathbb{R}^d$ are constraint sets. Problem (6.2.10) arises in the dual form soft-margin kernelized support vector machine classifier [54] in which $C_1$ is a box constraint and $C_2$ is a linear constraint. It also arises in portfolio optimization problems in which $C_1$ is a single linear inequality constraint and $C_2$ is the standard simplex. See Sections 6.5.3 and 6.5.4 for more details.

Define the smooth function $h(x) = \frac{1}{2} \langle Qx, x \rangle + \langle c, x \rangle$ and the nonsmooth indicator functions $g(x) = \iota_{C_1}(x)$ (which is 0 on $C_1$ and $\infty$ elsewhere) and $f := \iota_{C_2}$. This splitting is particularly nice because $\nabla h(x) = Qx + c$ is simple whereas the proximal operator of $h$ requires a matrix inversion, which is expensive for large scale problems. The proximal operators of $f$ and $g$ are just the projections onto $C_1$ and $C_2$, respectively.

Algorithm 6.2.4 (for problem (6.2.10)). Set an arbitrary $z^0$. For $k = 0, 1, \ldots$, iterate
1. get $w^k = P_{c_1}(z^k)$;

2. get $x^k = P_{c_2}(2w^k - z^k - \gamma(Qx + c))$;

3. get $z^{k+1} = z^k + x^k - w^k$.

The algorithm also works in the infinite dimension.

### 6.2.3 Simplest 3-block extension of ADMM

The 3-block monotropic program has the form

$$\begin{align*}
\text{minimize } & f_1(x_1) + f_2(x_2) + f_3(x_3) \\
\text{subject to } & L_1x_1 + L_2x_3 + L_3x_3 = b,
\end{align*}$$

where $\mathcal{H}_1, \ldots, \mathcal{H}_4$ are Hilbert spaces, the vector $b \in \mathcal{H}_4$ is given and for $i = 1, 2, 3$, the functions $f_i : \mathcal{H}_i \to (-\infty, \infty]$ are proper closed convex functions, and $L_i : \mathcal{H}_i \to \mathcal{H}_4$ are linear mappings. As usual, any constraint $x_i \in \mathcal{C}_i$ can be enforced through an indicator function $\iota_{\mathcal{C}_i}(x)$ and incorporated in $f_i$. We assume that $f_1$ is $\mu$-strongly convex where $\mu > 0$.

A new 3-block ADMM algorithm is obtained by applying Algorithm 6.1.1 to the dual formulation of (6.2.11) and rewriting the resulting algorithm using the original functions in (6.2.11). Let $f^*$ denote the convex conjugate of a function $f$, and let

$$
\begin{align*}
d_1(w) &:= f_1^*(L_1^*w), \\
d_2(w) &:= f_2^*(L_2^*w), \\
d_3(w) &:= f_3^*(L_3^*w) - \langle w, b \rangle.
\end{align*}
$$

The dual problem of (6.2.11) is

$$\min_w d_1(w) + d_2(w) + d_3(w).$$

Since $f_1$ is $\mu$-strongly convex, $d_1$ is $(\|L_1\|^2/\mu)$-Lipschitz continuous and, hence, the problem (6.2.12) is a special case of (6.2.4). We can adapt Algorithm 6.2.2 to (6.2.12) to get:
Algorithm 6.2.5 (for problem (6.2.12)). Set an arbitrary \( z^0 \) and stepsize \( \gamma \in (0, 2\mu/\|L_1\|^2) \).

For \( k = 0, 1, \ldots \), iterate

1. get \( w^k = \text{prox}_{\gamma d_3}(z^k) \);

2. get \( z^{k+\frac{1}{2}} = 2w^k - z^k - \gamma \nabla d_1(w^k) \);

3. get \( z^{k+1} = z^k + \text{prox}_{\gamma d_2}(z^{k+\frac{1}{2}}) - w^k \).

Throughout the rest of this section, we make the following assumption:

**Assumption 6.2.1.** For \( i = 1, 2, 3 \), the following differentiation rule holds:

\[
\partial d_{f_i}(x) = L_i^* \circ (\partial f_i^*) \circ L_i.
\]

See [11, Theorem 16.37] for conditions that imply this identity, of which the weakest are \( 0 \in \text{sri}(\text{range}(L_i^*) - \text{dom}(f_i^*)) \) where sri is the strong relative interior of a convex set. Notice that \( \text{dom}(f_i^*) = \mathcal{H} \) because \( f_1 \) is differentiable and thus, this assumption always holds. This assumption may seem strong, but it is standard in the analysis of ADMM because it implies the dual proximal operator identities in (2.8.4) and the identities in Proposition 6.2.1.

The following well-known proposition helps implement Algorithm 6.2.5 using the original objective functions instead of the dual functions \( d_i \).

**Proposition 6.2.1.** Let \( f \) be a closed proper convex function and let \( d(w) := f^*(A^*w) - \langle w, c \rangle \).

1. Let \( c = 0 \). Any \( x' \in \arg\min_x f(x) - \langle w, Ax \rangle \) obeys \( Ax' \in \partial d(w) \). If \( f \) is strictly convex, then \( Ax' = \nabla d(w) \).

2. Any \( x'' \in \arg\min_x f(x) + \frac{\gamma}{2}\|Ax - c - (1/\gamma)y\|^2 \) obeys \( Ax'' - c \in \partial d(\text{prox}_{\gamma d}(y)) \) and \( \text{prox}_{\gamma d}(y) = y - \gamma(Ax'' - c) \).

(We use “\( \in \)” with “\( \arg\min \)” since the minimizers are not unique in general.)
For notational simplicity, let
\[ s_\gamma(x_1, x_2, x_3, w) := L_1 x_1 + L_2 x_2 + L_3 x_3 - b - \frac{1}{\gamma} w. \]

By Proposition 6.2.1 and algebraic manipulation, we derive the following algorithm from Algorithm 6.2.5.

**Algorithm 6.2.6 (3-block ADMM).** Set an arbitrary \( w^0 \) and \( x^0_3 \), as well as stepsize \( \gamma \in (0, \frac{2\mu}{\|L_1\|^2}) \). For \( k = 0, 1, \ldots \), iterate

1. get \( x_1^{k+1} = \arg \min_{x_1} f_1(x_1) - \langle w^k, L_1 x_1 \rangle \);
2. get \( x_2^{k+1} \in \arg \min_{x_2} f_2(x_2) + \frac{\gamma}{2} \|s(x_1^{k+1}, x_2, x_3^k)\|^2; \)
3. get \( x_3^{k+1} \in \arg \min_{x_3} f_3(x_3) + \frac{\gamma}{2} \|s(x_1^{k+1}, x_2^{k+1}, x_3)\|^2; \)
4. get \( w^{k+1} = w^k - \gamma (L_1 x_1^{k+1} + L_2 x_2^{k+1} + L_3 x_3^{k+1} - b) \).

Note that Step 1 does not involve a quadratic penalty term, and it returns a unique solution because \( f_1 \) is strongly convex. In contrast, Steps 2 and 3 involve quadratic penalty terms and may have multiple solutions (though the products \( L_2 x_2^{k+1} \) and \( L_3 x_3^{k+1} \) are still unique.)

**Proposition 6.2.2.** If the initial points of Algorithms 6.2.5 and 6.2.6 satisfy \( z^0 = w^0 + \gamma (L_3 x_3^0 - b) \), then the two algorithms give the same sequence \( \{w^k\}_{k \geq 0} \).

The proposition is a well-known result based on Proposition 6.2.1 and algebraic manipulations; the interested reader is referred to Proposition 2.8.1 of Chapter 2. The convergence of Algorithm 6.2.6 is given in the following theorem.

**Theorem 6.2.1.** Let \( H_1, \ldots, H_4 \) be Hilbert spaces, \( f_i : H_i \to H_4 \) be proper closed convex functions, \( i = 1, 2, 3 \), and assume that \( f_1 \) is \( \mu \)-strongly convex. Suppose that the set \( S^* \) of the saddle-point solutions \((x_1, x_2, x_3, w) \in H_1 \times \cdots \times H_4\) to (6.2.11) is nonempty. Let \( \rho = \|L_1\|^2/\mu > 0 \) and pick \( \gamma \) satisfying

\[ 0 < \gamma < \frac{2}{\rho} \quad (6.2.13) \]
Then the sequences \( \{w^k\}_{k \geq 0}, \{L_2 x_2^k\}_{k \geq 0}, \) and \( \{L_3 x_3^k\}_{k \geq 0} \) of Algorithm 6.2.6 converge weakly to \( w^*, L_2 x_2^*, \) and \( L_3 x_3^* \), and \( \{x_1^k\}_{k \geq 0} \) converges strongly to \( x_1^* \), for some \( (w^*, x_1^*, x_2^*, x_3^*) \in S^* \).

**Remark 6.2.1.** Note that it is possible to replace Step 4 of Algorithm 6.2.6 with the update rule

\[
w^{k+1} = w^k - \alpha \gamma (L_1 x_1^{k+1} + L_2 x_2^{k+1} + L_3 x_3^{k+1} - b)
\]

where \( \alpha \in (0, \bar{\alpha}) \) and

\[
\bar{\alpha} = (2(1 - \rho \gamma))^{-1} \left( 1 - 2 \rho \gamma + \sqrt{(1 - 2 \rho \gamma)^2 + 4(1 - \rho \gamma)} \right) > 1,
\]

for \( \rho = \|L_1\|^2/\mu \). We do not pursue this generalization here due to lack of space.

Algorithm 6.2.6 generalizes several other algorithms of the alternating direction type.

**Proposition 6.2.3.**

1. Tseng’s alternating minimization algorithm is a special case of Algorithm 6.2.6 if the \( x_3 \)-block vanishes.

2. The (standard) ADMM is a special case of Algorithm 6.2.6 if the \( x_1 \)-block vanishes.

3. The augmented Lagrangian method (i.e., the method of multipliers) is a special case of Algorithm 6.2.6 if the \( x_1 \) - and \( x_2 \)-blocks vanish.

4. The Uzawa (dual gradient ascent) algorithm is a special case of Algorithm 6.2.6 if the \( x_2 \)- and \( x_3 \)-blocks vanish.

Recently, it was shown that the direct extension of ADMM to three blocks does not converge [38]. Compared to the recent work [33, 39, 70, 81, 84] on convergent 3-block extensions of ADMM, Algorithm 6.2.6 is the simplest and works under the weakest assumption. The first subproblem in Algorithm 6.2.6 does not involve \( L_2 \) or \( L_3 \), so it is simpler than the typical ADMM subproblem. While \( f_1 \) needs to be strongly convex, no additional assumptions on \( f_2, f_3 \) and \( L_1, L_2, L_3 \) are required for the extension. In comparison, [70] assume that \( f_1, f_2, f_3 \) are strongly convex functions. The condition is relaxed to two strongly convex functions in [39, 84] while [39] also needs \( L_1 \) to have full column rank. The papers [81, 33] further reduce the condition to one strongly convex function, and [81]
uses proximal terms in all the three subproblems and assumes some positive definitiveness conditions, and [33] assumes full column rankness on matrices $L_2$ and $L_3$. A variety of convergence rates are established in these papers. It is worth noting that the conditions assumed by the other ADMM extensions, beyond the strong convexity of $f_1$, are not sufficient for linear convergence, so in theory they do not necessarily convergence faster. In fact, some of the papers use additional conditions in order to prove linear convergence.

### 6.2.3.1 An $m$-block ADMM with $(m-2)$ strongly convex objective functions

There is a great benefit for not having a quadratic penalty term in Step 1 of Algorithm 6.2.6. When $f_1(x_1)$ is separable, Step 1 decomposes to independent sub-steps. Consider the extended monotropic program

$$\begin{align*}
\text{minimize} & \quad \bar{f}_1(\bar{x}_1) + \bar{f}_2(\bar{x}_2) + \cdots + \bar{f}_m(\bar{x}_m) \\
\text{subject to} & \quad \bar{L}_1\bar{x}_1 + \bar{L}_2\bar{x}_2 + \cdots + \bar{L}_m\bar{x}_m = b,
\end{align*}$$

where $\bar{f}_1, \ldots, \bar{f}_{m-2}$ are strongly convex and $\bar{f}_{m-1}, \bar{f}_m$ are convex (but not necessarily strongly convex.) Problem (6.2.14) is a special case of problem (6.2.11) if we group the first $m-2$ blocks. Specifically, we let $f_1(x_1) := \bar{f}_1(\bar{x}_1) + \cdots + \bar{f}_{m-2}(\bar{x}_{m-2})$, $f_2(x_2) := \bar{f}_{m-1}(\bar{x}_{m-1})$, $f_3(x_3) := \bar{f}_m(\bar{x}_m)$, and define $x_1, x_2, x_3, L_1, L_2, L_3$ in obvious ways. Define $\bar{s}_*(x_1, x_2, x_3, w) := \bar{L}_1\bar{x}_1 + \bar{L}_2\bar{x}_2 + \cdots + \bar{L}_m\bar{x}_m - b - \frac{1}{\gamma}w$. Then, it is straightforward to adapt Algorithm 6.2.6 for problem (6.2.14) as:

**Algorithm 6.2.7** (m-block ADMM). Set an arbitrary $w^0$ and $\bar{x}_m^0$, and stepsize $\gamma \in (0, \min\{2\|L_i\|/\mu_i \mid i = 1, \ldots, m-2\})$. For $k = 0, 1, \ldots$, iterate

1. get $\bar{x}_i^{k+1} = \arg\min_{x_i} \bar{f}_i(\bar{x}_i) - \langle w^k, \bar{L}_i\bar{x}_i \rangle$ for $i = 1, 2, \ldots, m-2$, in parallel;

2. get $\bar{x}^{k+1}_{m-1} \in \arg\min_{\bar{x}_{m-1}} \bar{f}_{m-1}(\bar{x}_{m-1}) + \frac{\gamma}{2}\|\bar{s}(\bar{x}_1^{k+1}, \ldots, \bar{x}_{m-2}^{k+1}, \bar{x}_{m-1}, \bar{x}_m^{k+1})\|^2$;

3. get $\bar{x}_m^{k+1} \in \arg\min_{\bar{x}_m} \bar{f}_m(\bar{x}_m) + \frac{\gamma}{2}\|\bar{s}(\bar{x}_1^{k+1}, \ldots, \bar{x}_{m-1}^{k+1}, \bar{x}_m)\|^2$.
4. get \( w^{k+1} = w^k - \gamma (\bar{L}_1 \bar{x}_1^{k+1} + \bar{L}_2 \bar{x}_2^{k+1} + \cdots + \bar{L}_m \bar{x}_m^{k+1} - b) \).

Note that we should make an assumption similar to the one in Assumption 6.2.1 for this form of Algorithm 6.2.7 to hold. All convergence properties of Algorithm 6.2.7 are identical to those of Algorithm 6.2.6.

6.2.4 Reducing the number of operators before splitting

Problems involving multiple operators can be reduced to fewer operators by applying grouping and lifting techniques. They allow Algorithm 6.1.1 and existing splitting schemes to handle four or more operators.

In general, two or more Lipschitz-differentiable functions (or cocoercive operators) can be grouped into one function (or one cocoercive operator, respectively). On the other hand, grouping nonsmooth functions with simple proximal maps (or monotone operators with simple resolvent maps) may lead to a much more difficult proximal map (or resolvent map, respectively). One resolution is lifting: to introduce dual and dummy variables and create fewer but “larger” operators. It comes with the cost that the introduced variables increase the problem size and may slow down convergence.

For example, we can reformulate Problem (6.1.1) in the form (which abuses the block matrix notation):

\[
0 \in \begin{bmatrix} B & I \\ -I & A^{-1} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} Cx \\ 0 \end{bmatrix} =: \bar{A} \begin{bmatrix} x \\ y \end{bmatrix} + \bar{C} \begin{bmatrix} x \\ y \end{bmatrix}.
\]  

(6.2.15)

Here we have introduced \( y \in Ax \), which is equivalent to \( x \in A^{-1}y \) or the second row of (6.2.15). Both the operators \( \bar{A} \) and \( \bar{C} \) are monotone, and the operator \( \bar{C} \) is cocoercive since \( C \) is so. Therefore, the problem (6.1.1) has been reduced to a monotone inclusion involving two “larger” operators. Under a special metric, applying the FBS iteration in [52] gives the following algorithm:
Algorithm 6.2.8 ([52]). Set an arbitrary $x^0, y^0$. Set stepsize parameters $\tau, \sigma$. For $k = 1, \ldots$, iterate:

1. get $x^k = J_{\tau B}(x^{k-1} - \tau C x^{k-1} - \tau y^{k-1})$;
2. get $y^k = J_{\sigma A^{-1}}(y^{k-1} + \sigma(2x^k - x^{k-1}))$  //comment: $J_{\sigma A^{-1}} = I - \sigma J_{\sigma^{-1}A} \circ (\sigma^{-1}I)$.

The lifting technique can be applied to the monotone inclusion problems with four or more operators together with Algorithm 6.1.1. Since Algorithm 6.1.1 handles three operators, it generally requires less lifting than previous algorithms. We re-iterate that FBS is a special case of our splitting, so Algorithm 6.2.8 is a special case of Algorithm 6.1.1 applied to (6.2.15) with a vanished $\bar{B}$.

Because both Algorithms 6.1.1 and 6.2.8 solve the problem (6.1.1), it is interesting to compare them. Note that one cannot obtain one algorithm from the other through algebraic manipulation. Both algorithms apply $J_A, J_B, C$ once every iteration. We managed to rewrite Algorithm 6.1.1 in the following equivalent form (see Appendix C.2 for a derivation) that is most similar to Algorithm 6.2.8 for the purpose of comparison:

Algorithm 6.2.9 (Algorithm 6.1.1 in an equivalent form). Set an arbitrary $x^0$ and $y^0$. For $k = 1, \ldots$, iterate:

1. get $x^k = J_{\gamma B}(x^{k-1} - \gamma C x^{k-1} - \gamma y^{k-1})$;
2. get $y^k = J_{\frac{1}{\gamma} A^{-1}}\left( y^{k-1} + \frac{1}{\gamma}(2x^k - x^{k-1}) + (C x^k - C x^{k-1}) \right)$  //comment: $J_{\sigma A^{-1}} = I - \sigma J_{\sigma^{-1}A} \circ (\sigma^{-1}I)$.

The difference between Algorithms 6.2.8 and 6.2.9 is the extra correction factor $C x^k - C x^{k-1}$. Without the correction factor, we cannot eliminate $y^k$ and express Algorithms 6.2.8 in the form of (6.1.4).
6.3 Convergence theory

In this section, we show that Problem (6.1.1) can be solved by iterating the operator $T$ defined in Equation (6.1.3): $T = I_H - J_{\gamma_B} + J_{\gamma_A} \circ (2J_{\gamma_B} - J_{\gamma_C})$.

In Figure 6.1, we depict the process of applying $T$ to a point $z \in H$. Lemma 6.3.1 defines the points in Figure 6.1.

**Lemma 6.3.1.** Let $z \in H$ and define points:

- $x_B^k := J_{\gamma_B}(z^k)$
- $z^k := 2x_B^k - z^k$
- $x_A^k := J_{\gamma_A}(z'')$
- $u_B^k := \gamma^{-1}(z^k - x_B^k) \in Bx_B^k$
- $z'' := z' - \gamma Cx_B^k$
- $u_A^k := \gamma^{-1}(z'' - x_A^k) \in Ax_A^k$

Then the following identities hold:

$$Tz^k - z^k = x_A^k - x_B^k = -\gamma(u_B^k + u_A^k + Cx_B^k) \quad \text{and} \quad Tz^k = x_A^k + \gamma u_B^k.$$ 

When $B = \partial g$, we let $\tilde{\nabla} g(x^k_g) := u_B^k \in \partial g(x^k_g)$. Likewise when $A = \partial f$, we let $\tilde{\nabla} f(x^k_f) := u_A^k \in \partial f(x^k_f)$.
Proof. Observe that $Tz^k = z^k + x_A^k - x_B^k$ by the definition of $T$ (Equation (6.1.3)). In addition, $Tz^k = x_A^k + z^k - x_B^k = x_A^k + \gamma u_B^k$. Finally, we have $x_A^k - x_B^k = 2x_B^k - z^k - \gamma u_A^k - \gamma C x_B^k - x_B^k = -\gamma(u_A^k + u_B^k + C x_B^k)$. 

The following proposition computes a fixed point identity for the operator $T$. It shows that we can recover a zero of $A + B + C$ from any fixed point $z^*$ of $T$ by computing $J_{\gamma B} z^*$. 

**Lemma 6.3.2 (Fixed-point encoding).** The following set equality holds

$$\text{zer}(A + B + C) = J_{\gamma B} (\text{Fix} T).$$

In addition, 

$$\text{Fix} T = \{ x + \gamma u \mid 0 \in (A + B + C)x, u \in (Bx) \cap (-Ax - Cx) \}.$$ 

The proof can be found in Appendix C.3.

The next lemma will help us establish the averaged coefficient of the operator $T$ in the next proposition. Note that in the lemma, if we let $W := 0$, $U := I_H - J_{\gamma B}$, and $T_1 := J_{\gamma A}$, the operator $S$ reduces to the DRS operator $I_H - J_{\gamma B} + J_{\gamma A} \circ (2J_{\gamma B} - I_H)$, which is known to be $1/2$-averaged.

**Lemma 6.3.3.** Let $S := U + T_1 \circ V$, where $U, T_1 : \mathcal{H} \to \mathcal{H}$ are both firmly nonexpansive and $V : \mathcal{H} \to \mathcal{H}$. Let $W = I - (2U + V)$. Then we have for all $z, w \in \mathcal{H}$:

$$\|Sz - Sw\|^2 \leq \|z - w\|^2 - \|(I_H - S)z - (I_H - S)w\|^2 - 2(T_1 \circ V z - T_1 \circ V w, W z - W w).$$ (6.3.1)

The proof can be found in Appendix C.3.

The following proposition will show that the operator $T$ is averaged. This proposition is crucial for proving the convergence of Algorithm 6.1.1.

**Proposition 6.3.1 (Averageness of $T$).** Suppose that $T_1, T_2 : \mathcal{H} \to \mathcal{H}$ are firmly nonexpansive and $C$ is $\beta$-cocoercive, $\beta > 0$. Let $\gamma \in (0, 2\beta)$. Then

$$T := I - T_2 + T_1 \circ (2T_2 - I_H - \gamma C \circ T_2)$$
is $\alpha$-averaged with coefficient $\alpha := \frac{2\beta}{\beta - \gamma} < 1$. In particular, the following inequality holds for all $z, w \in \mathcal{H}$

$$\|Tz - Tw\|^2 \leq \|z - w\|^2 - \frac{(1 - \alpha)}{\alpha} \|(I_{\mathcal{H}} - T)z - (I_{\mathcal{H}} - T)w\|^2.$$  \hfill (6.3.2)

**Proof.** To apply Lemma 6.3.3, we let $U := I_{\mathcal{H}} - T_2$, $V := 2T_2 - I_{\mathcal{H}} - \gamma C \circ T_2$, and $W := \gamma C \circ T_2$. Note that $U$ is firmly nonexpansive (because $T_2$ is), and we have $W = I_{\mathcal{H}} - (2U + V)$. Let $S := T = I_{\mathcal{H}} - T_2 + T_1 \circ V$. We evaluate the inner product in (6.3.1) as follows:

$$-2\langle T_1 \circ Vz - T_1 \circ Vw, Wz - Ww \rangle$$

$$= 2\langle (I_{\mathcal{H}} - T)z - (I_{\mathcal{H}} - T)w, \gamma C \circ T_2 z - \gamma C \circ T_2 w \rangle$$

$$-2\langle T_2 z - T_2 w, \gamma C \circ T_2 z - \gamma C \circ T_2 w \rangle$$

$$\leq \varepsilon \|(I_{\mathcal{H}} - T)z - (I_{\mathcal{H}} - T)w\|^2 + \frac{\gamma^2}{\varepsilon} \|(C \circ T_2 z - C \circ T_2 w)\|^2$$

$$-2\gamma \beta \|(C \circ T_2 z - C \circ T_2 w)\|^2$$

$$= \varepsilon \|(I_{\mathcal{H}} - T)z - (I_{\mathcal{H}} - T)w\|^2 - \gamma (2\beta - \gamma/\varepsilon) \|(C \circ T_2 z - C \circ T_2 w)\|^2$$

where the inequality follows from Young’s inequality with any $\varepsilon > 0$ and that $C$ is $\beta$-cocoercive. We set

$$\varepsilon := \gamma/2\beta < 1$$

so that the coefficient $\gamma(2\beta - \gamma/\varepsilon) = 0$. Now applying Lemma 6.3.3 and using $S = T$, we obtain

$$\|Tz - Tw\|^2 \leq \|z - w\|^2 - (1 - \varepsilon) \|(I_{\mathcal{H}} - T)z - (I_{\mathcal{H}} - T)w\|^2,$$

which is identical to (6.3.2) under our definition of $\alpha$. \hfill $\square$

**Remark 6.3.1.** It is easy to slightly strengthen the inequality (6.3.2) as follows: For any $\bar{\varepsilon} \in (0, 1)$ and $\bar{\gamma} \in (0, 2\beta \bar{\varepsilon})$, let $\bar{\alpha} := 1/(2 - \bar{\varepsilon}) < 1$. Then the following holds for all $z, w \in \mathcal{H}$:

$$\|Tz - Tw\|^2 \leq \|z - w\|^2 - \frac{(1 - \bar{\alpha})}{\bar{\alpha}} \|(I_{\mathcal{H}} - T)z - (I_{\mathcal{H}} - T)w\|^2$$

$$- \bar{\gamma} \left(\frac{2\beta - \frac{\bar{\gamma}}{\bar{\varepsilon}}}{\bar{\varepsilon}}\right) \|(C \circ T_2(z) - C \circ T_2(w))\|^2.$$

\hfill (6.3.3)
Remark 6.3.2. When \( C = 0 \), the mapping in Equation (6.1.5) reduces to \( S = \text{refl}_{\gamma A} \circ \text{refl}_{\gamma B} \), which is nonexpansive because it is the composition of nonexpansive maps. Thus, \( T = (1/2)I_H + (1/2)S \) is firmly nonexpansive by definition. However, when \( C \neq 0 \), the mapping \( S \) in (6.1.5) is no longer nonexpansive. The mapping \( 2T_2 - I_H - \gamma C \), which is a part of \( S \), can be expansive. Indeed, consider the following example: Let \( H = \mathbb{R}^2 \), let \( B = \partial \iota \{ (x_1,0) | x_1 \in \mathbb{R} \} \) be the normal cone of the \( x_1 \) axis, and let \( C = \nabla ((1/2)\|x_1 + x_2\|^2) = (x_1 + x_2, x_1 + x_2) \). In particular, \( T_2(x_1, x_2) = J_{\gamma B}(x_1, x_2) = (x_1, 0) \) for all \((x_1, x_2) \in \mathbb{R}^2 \) and \( \gamma > 0 \). Then the point 0 is a fixed point of \( R = 2T_2 - I_H - \gamma C \circ T_2 \), and

\[
R(1,1) = (1, -1) - \gamma C(1, 0) = (1 - \gamma, -1 - \gamma).
\]

Therefore, \( \|R(1,1) - R(0,0)\| = \sqrt{2 + 2\gamma^2} > \sqrt{2} = \|(1,1) - (0,0)\| \) for all \( \gamma > 0 \).

Remark 6.3.3. When \( B = 0 \), the averaged parameter \( \alpha = 2\beta/(4\beta - \gamma) \) in Proposition 6.3.1 reduces to the best (i.e., smallest) known averaged coefficient for the forward-backward splitting algorithm [50, Proposition 2.4].

We are now ready to prove convergence of Algorithm 6.1.1.

Theorem 6.3.1 (Main convergence theorem). Suppose that \( \text{Fix } T \neq \emptyset \). Set a stepsize \( \gamma \in (0, 2\beta \varepsilon) \), where \( \varepsilon \in (0, 1) \). Set \( (\lambda_j)_{j \geq 0} \subseteq (0, 1/\alpha) \) as a sequence of relaxation parameters, where \( \alpha = 1/(2 - \varepsilon) < 2\beta/(4\beta - \gamma) \), such that for \( \tau_k := (1 - \lambda_k/\alpha)\lambda_k/\alpha \) we have \( \sum_{i=0}^{\infty} \tau_i = \infty \). Pick any start point \( z^0 \in H \). Let \( (z^j)_{j \geq 0} \) be generated by Algorithm 6.1.1, i.e., the following iteration: for all \( k \geq 0 \),

\[
z^{k+1} = z^k + \lambda_k(Tz^k - z^k).
\]

Then the following hold

1. Let \( z^* \in \text{Fix } T \). Then \( (\|z^j - z^*\|)_{j \geq 0} \) is monotonically decreasing.

2. The sequence \( (\|Tz^j - z^j\|)_{j \geq 0} \) is monotonically decreasing and converges to 0.

3. The sequence \( (z^j)_{j \geq 0} \) weakly converges to a fixed point of \( T \).
4. Let $x^* \in \text{zer}(A + B + C)$. Suppose that $\inf_{j \geq 0} \lambda_j > 0$. Then the following sum is finite:

$$
\sum_{i=0}^{\infty} \lambda_i \|Cx_B^i - Cx^*\|^2 \leq \frac{1}{\gamma(2\beta - \gamma/\varepsilon)} \|z^0 - z^*\|^2
$$

In particular, $(Cx_B^j)_{j \geq 0}$ converges strongly to $Cx^*$.

5. Suppose that $\inf_{j \geq 0} \lambda_j > 0$ and let $z^*$ be the weak sequential limit of $(z^j)_{j \geq 0}$. Then the sequence $(J_{\gamma B}(z^j))_{j \geq 0}$ weakly converges to $J_{\gamma B}(z^*) \in \text{zer}(A + B + C)$.

6. Suppose that $\inf_{j \geq 0} \lambda_j > 0$ and let $z^*$ be the weak sequential limit of $(z^j)_{j \geq 0}$. Then the sequence $(J_{\gamma A} \circ (2J_{\gamma B} - I_H - \gamma C \circ J_{\gamma B})(z^j))_{j \geq 0}$ weakly converges to $J_{\gamma B}(z^*) \in \text{zer}(A + B + C)$.

7. Suppose that $\tau := \inf_{j \geq 0} \tau_j > 0$. For all $k \geq 0$, the following convergence rates hold:

$$
\|Tz^k - z^k\|^2 \leq \frac{\|z^0 - z^*\|^2}{\tau(k + 1)} \quad \text{and} \quad \|Tz^k - z^k\|^2 = o\left(\frac{1}{k + 1}\right)
$$

for any point $z^* \in \text{Fix}(T)$.

8. Let $z^*$ be the weak sequential limit of $(z^j)_{j \geq 0}$. The sequences $(J_{\gamma B}(z^j))_{j \geq 0}$ and $(J_{\gamma A} \circ (2J_{\gamma B} - I_H - \gamma C \circ J_{\gamma B})(z^j))_{j \geq 0}$ converge strongly to a point in $\text{zer}(A + B + C)$ whenever any of the following holds:

(a) $A$ is uniformly monotone\(^3\) on every nonempty bounded subset of $\text{dom}(A)$;

(b) $B$ is uniformly monotone on every nonempty bounded subset of $\text{dom}(B)$;

(c) $C$ is demiregular at every point $x \in \text{zer}(A + B + C)$.\(^4\)

---

\(^3\)A mapping $A$ is uniformly monotone if there exists increasing function $\phi : \mathbb{R}_+ \to [0, +\infty]$ such that $\phi(0) = 0$ and for any $u \in Ax$ and $v \in Ay$, $\langle x - y, u - v \rangle \geq \phi(\|x - y\|)$. If $\phi \equiv \beta(\cdot)^2 > 0$, then the mapping $A$ is strongly monotone. If a proper function $f$ is uniformly (strongly) convex, then $\partial f$ is uniformly (strongly, resp.) monotone.

\(^4\)A mapping $C$ is demiregular at $x \in \text{dom}(C)$ if for all $u \in Cx$ and all sequences $(x^k, u^k) \in \text{gra}(C)$ with $x^k \rightarrow x$ and $u^k \rightarrow u$, we have $x^k \rightarrow x$. 

237
Proof. Part 1: Fix \( k \geq 0 \). Observe that
\[
\|z^{k+1} - z^*\|^2 = \|(1 - \lambda_k)(z^k - z^*) + \lambda_k(Tz^k - z^*)\|^2
\]
\[
= (1 - \lambda_k)\|z^k - z^*\|^2 + \lambda_k\|Tz^k - z^*\|^2 - \lambda_k(1 - \lambda_k)\|Tz^k - z^k\|^2
\]
(6.3.4)
by Corollary [11, Corollary 2.14]. In addition, from Equation (6.3.3), we have
\[
\|Tz^k - z^*\|^2 \leq \|z^k - z^*\|^2 - \frac{1 - \alpha}{\alpha}\|Tz^k - z^k\|^2
\]
\[
- \gamma \left(2\beta - \frac{\gamma}{\varepsilon}\right) \|C \circ T(z^k) - C \circ T(z^*)\|^2.
\]
Therefore, the monotonicity follows by combining the above two equations and using the simplification
\[
\tau_k = \lambda_k(1 - \lambda_k) + \frac{\lambda_k(1 - \alpha)}{\alpha}
\]
to get
\[
\|z^{k+1} - z^*\|^2 + \tau_k\|Tz^k - z^k\|^2 + \gamma\lambda_k \left(2\beta - \frac{\gamma}{\varepsilon}\right) \|Cx^k_B - CJ\gamma B(z^*)\|^2 \leq \|z^k - z^*\|^2.
\]

Part 2: This follows from [11, Proposition 5.15(ii)].

Part 3: This follows from [11, Proposition 5.15(iii)].

Part 4: The inequality follows by summing the last inequality derived in Part 1. The convergence of \((Cx^j_B)_{j \geq 0}\) follows because \(\inf_{j \geq 0} \lambda_j > 0\) and the sum is finite.

Part 5: Recall the notation from Lemma 6.3.1: set \(x^k_B = J_{\gamma B}(z^k)\), \(x^k_A = J_{\gamma A}(2x^k_B - z^k - \gamma Cx^k_B)\), \(u^k_B = (1/\gamma)(z^k - x^k_B) \in Bx^k_B\), and \(u^k_A = (1/\gamma)(2x^k_B - z^k - \gamma Cx^k_B - x^k_A) \in Ax^k_A\).

Since \(\|x^k_B - J_{\gamma B}(z^*)\| = \|J_{\gamma B}(z^k) - J_{\gamma B}(z^*)\| \leq \|z^k - z^*\| \leq \|z^0 - z^*\|\), \(\forall k \geq 0\), the sequence \((x^j_B)_{j \geq 0}\) is bounded and has a weak sequential cluster point \(\overline{x}\). Let \(x^{k_j}_B \rightharpoonup \overline{x}\) as \(j \to \infty\) for index subsequence \((k_j)_{j \geq 0}\).

Let \(x^* \in \text{zer}(A + B + C)\). Because \(C\) is maximal monotone, \(Cx^k_B \to Cx^*\), and \(x^k_B \rightharpoonup \overline{x}\), it follows by the weak-to-strong sequential closedness of \(C\) that \(C\overline{x} = Cx^*\) [11, Proposition
20.33(ii)] and thus \( Cx_B^{k_j} \to C\overline{x} \). Because \( x_A^k - x_B^k = Tz^k - z^k \to 0 \) as \( k \to \infty \) by Part 2 and Lemma 6.3.1, it follows that 

\[
x_B^{k_j} \to \overline{x}, \quad x_A^{k_j} \to \overline{x}, \quad Cx_B^{k_j} \to C\overline{x}, \quad u_B^{k_j} \to \frac{1}{\gamma} (z^* - \overline{x}), \quad \text{and} \quad u_A^{k_j} \to \frac{1}{\gamma} (\overline{x} - z^* - \gamma C\overline{x})
\]
as \( j \to \infty \).

Thus, [11, Proposition 25.5] applied to \((x_A^{k_j}, u_A^{k_j}) \in \text{gra} \, A, (x_B^{k_j}, u_B^{k_j}) \in B, \) and \((x_B^{k_j}, Cx_B^{k_j}) \in C\) shows that \( \overline{x} \in \text{zer}(A + B + C) \), \( z^* - \overline{x} \in \gamma B\overline{x} \), and \( \overline{x} - z^* - \gamma C\overline{x} \in \gamma A\overline{x} \). Hence, as 

\( \overline{x} = J_{\gamma B}(z^*) \) is unique, \( \overline{x} \) is the unique weak sequential cluster point of \((x_B^j)_{j \geq 0}\). Therefore, 

\((x_B^j)_{j \geq 0}\) converges weakly to \( J_{\gamma B}(z^*) \) by [11, Lemma 2.38].

Part 6: Assume the notation of Part 5. We shall show \( x_A^k \to J_{\gamma B}(z^*) \). This follows because \( x_A^k - x_B^k = Tz^k - z^k \to 0 \) as \( k \to \infty \) and \( x_B^k \to J_{\gamma B}(z^*) \).

Part 7: The result follows from Theorem 2.3.1 of Chapter 2.

Part 8: Assume the notation of Part 5 and let \( x^* = J_{\gamma B}(z^*) \), \( u_B^* = (1/\gamma)(z^* - x^*) \in Bx^* \), and \( u_A^* = (1/\gamma)(z^* - x^*) - Cx^* \). Now we move to the subcases.

Part 8a: Because \( B + C \) is monotone and \((x_B^k, u_B^k) \in B\), we have \( (x_B^k - x^*, u_B^k + Cx_B^k - (u_B^* + Cx_B^*)) \geq 0 \) for all \( k \geq 0 \). Consider the bounded set \( S = \{x^*\} \cup \{x_A^j | j \geq 0\} \). Then there exists an increasing function \( \phi_A : \mathbb{R}_+ \to [0, \infty] \) that vanishes only at 0 such that 

\[
\gamma \phi_A(||x_A^k - x^*||) \leq \gamma \langle x_A^k - x^*, u_A^k - u_A^* \rangle + \gamma \langle x_B^k - x^*, u_B^k + Cx_B^k - (u_B^* + Cx_B^*) \rangle
\]

\[
= \gamma \langle x_A^k - x_B^k, u_A^k - u_A^* \rangle + \gamma \langle x_B^k - x^*, u_B^k - u_B^* \rangle
\]

\[
+ \gamma \langle x_B^k - x^*, u_B^k + Cx_B^k - (u_B^* + Cx_B^*) \rangle
\]

\[
= \gamma \langle x_A^k - x_B^k, u_A^k - u_A^* \rangle + \gamma \langle x_B^k - x^*, u_A^k + u_B^k + Cx_B^k \rangle
\]

\[
= \langle x_B^k - x_A^k, x_B^k - \gamma u_A^k - (x^* - \gamma u_A^*) \rangle
\]

\[
= \langle x_B^k - x_A^k, z^k - z^* \rangle + \gamma \langle x_B^k - x_A^k, Cx_B^k - Cx^* \rangle \to 0 \quad \text{as} \quad k \to \infty
\]

where the convergence to 0 follows because \( x_B^k - x_A^k = z^k - Tz^k \to 0, z^k \to z^*, \) and \( Cx_B^k \to Cx^* \) as \( k \to \infty \). Furthermore, \( x_B^k \to x^* \) because \( x_A^k - x_B^k \to 0 \) as \( k \to \infty \).

Part 8b: Because \( A \) is monotone, we have \( (x_A^k - x^*, u_A^k - u_A^*) \geq 0 \) for all \( k \geq 0 \). In addition, note that \( B + C \) is also uniformly monotone on all bounded sets. Consider
the bounded set $S = \{x^*\} \cup \{x^j_B \mid j \geq 0\}$. Then there exists an increasing function
$\phi_B : \mathbb{R}_+ \to [0, \infty]$ that vanishes only at 0 such that

$$\phi_B(\|x^k_B - x^*\|) \leq \gamma \langle x^k_B - x^*, u^k_B - u_A^* \rangle + \gamma \langle x^k_B - x^*, u_B^k + Cx_B^k - (u_B^* + Cx^*_B) \rangle \to 0 \text{ as } k \to \infty$$

by the argument in Part 8a. Therefore, $x^k_B \to x^*$ strongly.

Part 8c: Note that $Cx^k_B \to Cx^*$ and $x^k_B \rightharpoonup x^*$. Therefore, $x^k_B \to x^*$ by the demiregularity of $C$. \hfill \Box

Remark 6.3.4. Theorem 6.3.1 can easily be extended to the summable error scenario, where for all $k \geq 0$, we have

$$z^{k+1} = z^k + \lambda_k(Tz^k - z^k + e_k)$$

for a sequence $(e_j)_{j \geq 0} \subseteq \mathcal{H}$ of errors that satisfy $\sum_{i=0}^{\infty} \lambda_k \|e_j\| < \infty$ (e.g., using [50, Proposition 3.4]). The result is straightforward and will only serve to complicate notation, so we omit this extension.

Remark 6.3.5. Note that by Theorem 2.3.1 of Chapter 2, the convergence rates for the fixed-point residual in Part 7 of Theorem 6.3.1 are sharp—even in the case of the variational Problem (6.1.2) with $h = 0$.

6.4 Convergence rates

In this section, we discuss the convergence rates Algorithm 6.1.1 under several different assumptions on the regularity of the problem. Section 6.1.2 contains a brief overview of all the convergence rates presented in this section. For readability, we now summarize all of the convergence results of this section, briefly indicate the proof structure, and place the formal proofs in the Appendix.
6.4.1 General rates

We establish our most general convergence rates for the following quantities: If \( z^* \) is a fixed point of \( T \), \( x^* = J_{\gamma B}(z^*) \), and \( x \in \mathcal{H} \), then let

\[
\kappa_1^k(\lambda, x) = \|z^k - x\|^2 - \|z^{k+1} - x\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z^k - z^{k+1}\|^2 + 2\gamma \langle z^k - z^{k+1}, Cx_B \rangle,
\]

\[
\kappa_2^k(\lambda, x^*) = \|z^k - z^\ast\|^2 - \|z^{k+1} - z^\ast\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z^k - z^{k+1}\|^2 + 2\gamma \langle z^k - z^{k+1}, Cx_B - Cx^\ast \rangle.
\]

In Theorems C.1, C.2, and C.3 we deduce the following convergence rates: For \( j \in \{1, 2\} \), \( x_1 = x \), \( x_2 = x^* \) and for all \( k \geq 0 \), we have

Nonergodic (Algorithm 6.1.1):

\[
\kappa_j^k(1, x_j) = o\left(\frac{1 + \|x_j\|}{\sqrt{k + 1}}\right);
\]

Ergodic (Algorithm 6.1.1 & (6.1.6)):

\[
\frac{1}{\sum_{i=0}^{k} \lambda_k} \sum_{i=0}^{k} \kappa_j^i(\lambda, x_j) = O\left(\frac{1 + \|x_j\|^2}{k + 1}\right);
\]

Ergodic (Algorithm 6.1.1 & (6.1.7)):

\[
\frac{2}{(k + 1)(k + 2)} \sum_{i=0}^{k} (i + 1)\kappa_j^i(\lambda, x_j) = O\left(\frac{1 + \|x_j\|^2}{k + 1}\right).
\]

It may be hard to see how these terms relate to the convergence of Algorithm 6.1.1. The key observation of Proposition C.3 shows that \( \kappa_1^k \) is an upper bound for a certain variational inequality associated to Problem (6.1.1) and that \( \kappa_2^k \) bounds the distance of the current iterate \( x^k \) (or its averaged variant \( x^k \) in Equations (6.1.6) and (6.1.7)) to the solution whenever one of the operators is strongly monotone.

The proofs of these convergence rates are straightforward, though technical. The nonergodic rates follow from an application of Part 7 of Theorem 6.3.1, which shows that \( \|z^{k+1} - z^k\|^2 = o(1/(k + 1)) \). The ergodic convergence rates follow from the alternating series properties of \( \kappa_j^k \) together with the summability of the gradient shown in Part 4 of Theorem 6.3.1.

241
6.4.2 Objective error and variational inequalities

In this section, we use the convergence rates of the upper and lower bounds derived in Theorems C.1, C.2, and C.3 to deduce convergence rates of function values and variational inequalities. All of the convergence rates have the following orders:

\[
\text{Nonergodic: } o\left(\frac{1}{\sqrt{k} + 1}\right) \quad \text{and} \quad \text{Ergodic: } O\left(\frac{1}{k + 1}\right).
\]

The convergence rates in this section generalize some of the known convergence rates provided in Chapters 2, 4, and 5 for Douglas-Rachford splitting, forward-Douglas-Rachford splitting, and the primal-dual forward-backward splitting, Douglas-Rachford splitting, and the proximal-point algorithms.

6.4.2.1 Nonergodic rates (Algorithm 6.1.1)

Suppose that \( A = \partial f + \overline{A}, \ B = \partial g + \overline{B} \) and \( C = \nabla h + \overline{C} \) where \( f, g \) and \( h \) are functions and \( \overline{A}, \overline{B} \) and \( \overline{C} \) are monotone operators. Whenever \( f \) and \( \overline{A} \) are Lipschitz continuous, the following convergence rate holds:

\[
f(x_B^k) + g(x_B^k) + h(x_B^k) - (f + g + h)(x) + \langle x_B^k - x, \overline{A}x_B^k + u_B^k + \overline{C}x_B^k \rangle = o\left(\frac{1 + \|x\|}{\sqrt{k + 1}}\right).
\]  

(6.4.1)

A more general rate holds when \( f \) and \( \overline{A} \) are not necessarily Lipschitz. See Corollaries C.3 and C.4 for the exact convergence statements.

Note that quantity on the left hand side of Equation (6.4.1) can be negative. The point \( x_B^k \) is a solution to the variational inequality problem if, and only if, the Equation (6.4.1) is negative for all \( x \in \mathcal{H} \), which is why we include the dependence on \( \|x\| \).

Notice that when the operators \( \overline{A}, \overline{B} \) and \( \overline{C} \) vanish and \( x = x^* \), the convergence rate in (6.4.1) reduces to the objective error of the function \( f + g + h \) at the point \( x_B^k \)

\[
f(x_B^k) + g(x_B^k) + h(x_B^k) - (f + g + h)(x^*) = o\left(\frac{1 + \|x^*\|}{\sqrt{k + 1}}\right).
\]  

(6.4.2)
and we deduce the rate $o(1/\sqrt{k+1})$ for our method. By Theorem 2.7.2 of Chapter 2, this rate is sharp.

Further nonergodic rates can be deduced whenever any $A$, $B$, or $C$ are $\mu_A$, $\mu_B$ and $\mu_C$-strongly monotone respectively. In particular, the following two rates hold for all $k \geq 0$ (Corollary C.9):

$$\mu_A\|x_A^k - x^*\|^2 + (\mu_B + \mu_C)\|x_B^k - x^*\|^2 = o\left(\frac{1}{\sqrt{k+1}}\right);$$

$$\min_{i=0,\ldots,k} \left\{ \mu_A\|x_A^i - x^*\|^2 + (\mu_B + \mu_C)\|x_B^i - x^*\|^2 \right\} = o\left(\frac{1}{k+1}\right).$$

### 6.4.2.2 Ergodic Rates

We use the same set up as Section 6.4.2, except we assume that $\overline{A}$ and $\overline{B}$ are skew linear mappings (i.e., $A^* = -A$ and $B^* = -B$) and $\overline{C} = 0$. If $(\overline{x}_B^j)_{j \geq 0}$ is generated as in Equation (6.1.6) or Equation (6.1.7) and $f$ is Lipschitz continuous, the following convergence rate holds:

$$f(\overline{x}_B^k) + g(\overline{x}_B^k) + h(\overline{x}_B^k) - (f + g + h)(x) + \langle \overline{x}_B^k - x, \overline{A}\overline{x}_B^k + \overline{B}\overline{x}_B^k \rangle = o\left(\frac{1 + \|x\|^2}{k+1}\right).$$

(6.4.3)

A more general rate holds when $f$ is not necessarily Lipschitz. See Corollaries C.5–for the exact convergence statements.

Further nonergodic rates can be deduced whenever any $A$, $B$, or $C$ are $\mu_A$, $\mu_B$ and $\mu_C$-strongly monotone respectively. In particular, the following two rates hold for all $k \geq 0$ Corollary C.9: Let $(\overline{x}_A^j)_{j \geq 0}$ and $(\overline{x}_B^j)_{j \geq 0}$ be generated by Algorithm 6.1.1 and Equations (6.1.6) or (6.1.7). Then

$$\mu_A\|\overline{x}_A^k - x^*\|^2 + (\mu_B + \mu_C)\|\overline{x}_B^k - x^*\|^2 = O\left(\frac{1}{k+1}\right).$$

### 6.4.3 Improving the objective error with Lipschitz differentiability

The worst case convergence rate $o(1/\sqrt{k+1})$ for objective error discussed in proved in Corollary C.3 is quite slow. Although averaging can improve the rate of convergence, this
technique does not necessarily translate into better practical performance as discussed in Section 6.1.3.1. We can deduce a better rate of convergence for the nonergodic iterate, whenever one of the functions \(f\) or \(g\) has a Lipschitz continuous derivative. In particular, if \(\nabla f\) exists and is Lipschitz, we show in Proposition C.4 that the objective error sequence \(((f + g + h)(x_j^B) - (f + g + h)(x^*))_{j \geq 0}\) is summable. From this, we immediately deduce Theorem C.5 the following rate: for all \(k \geq 0\), we have
\[
\min_{i=0, \ldots, k} \{ (f + g + h)(x_B^i) - (f + g + h)(x^*) \} = o \left( \frac{1}{k+1} \right).
\]
A similar result holds for the objective error sequence \(((f + g + h)(x_A^j) - (f + g + h)(x^*))_{j \geq 0}\) when the function \(g\) is Lipschitz differentiable. Thus, when \(f\) or \(g\) is sufficiently regular, the convergence rate of the nonergodic iterate is actually faster than the convergence rate for the ergodic iterate, which motivates its use in practice.

### 6.4.4 Linear convergence

Whenever \(A, B\) and \(C\) are sufficiently regular, we can show that the operator \(T\) is strictly contractive towards the fixed point set. In particular, Algorithm 6.1.1 converges linearly whenever
\[
(\mu_A + \mu_B + \mu_C)(1/L_A + 1/L_B) > 0
\]
where \(L_A\) and \(L_B\) are the Lipschitz constants of \(A\) and \(B\) respectively and \(A, B,\) or \(C\) are \(\mu_A,\ \mu_B\) and \(\mu_C\)-strongly monotone respectively (where we allow the \(L_A = L_B = \mu_A = \mu_B = \mu_C = 0\)).

Note that this linear convergence result is the best we can expect in some sense. Indeed, even if \(\mu_C\) and \(\mu_A\) are strongly monotone, Algorithm 6.1.1 will not necessarily converge linearly. Section C.4.6 we provide an example such that
\[
\mu_A \mu_C > 0, \text{ but } \|z^k - z^*\| \text{ converges arbitrarily slowly to } 0.
\]
6.4.5 Convergence rates for multi-block ADMM

All of the results in this section imply convergence rates for Algorithm 6.2.6, which is applied to the dual objective in Problem (6.2.12). Using the techniques of Chapter 2 and 3, we can easily derive convergence rates of the primal objective in Problem (6.2.11). We do not pursue these results in this chapter due to lack of space.

6.5 Numerical results

In this section, we present some numerical examples of Algorithm 6.1.1. We emphasize that to keep our implementations simple, we did not attempt to optimize the codes or their parameters for best performance. We also did not attempt to seriously evaluate the prediction ability of the models we tested, which is beyond the scope of this chapter. Our Matlab codes will be released online on the authors’ websites. All tests were run on a PC with 32GB memory and an Intel i5-3570 CPU with Ubuntu 12.04 and Matlab R2011b installed.

6.5.1 Image inpainting with texture completion

This section presents the results of applying Problem (6.2.8) to the color images of a building, parts of which are manually occluded with white colors. See Figure 6.2. The images have a 517 × 493 resolution and three color channels. At each iteration of Algorithm 6.1.1, the SVDs of two matrices of sizes 517 × 1479 and 1551 × 493 consume most of the computing time. However, it took less 150 iterations to return good recoveries.

\footnote{We are grateful of Professor Ji Liu for sharing his data in [86] with us.}
Figure 6.2: Images recovered by solving the tensor completion Problem (6.2.8) using Algorithm 6.1.1 for two different types of occlusions.
6.5.2 Matrix completion for movie recommendations

In this section, we apply Problem (6.2.9) to a movie recommendation dataset. In this example, each row of $X_0 \in \mathbb{R}^{m \times n}$ corresponds to a user and each column corresponds to a movie, and for all $i = 1, \ldots, m$ and $j = 1, \ldots, m$, the matrix entry $(X_0)_{ij}$ is the ranking that user $i$ gave to movie $j$.

We use the MovieLens-1M [1] dataset for evaluation. This dataset consists of 1000209 observations of the matrix $X_0 \in \mathbb{R}^{6040 \times 3952}$. We plot our numerical results in Figure 6.3. In our code we set $l = 0$, $u = 5$ and solved the problem with different choices of $\mu$ in order to achieve solutions of desired rank. In Figure 6.3c we plot the root mean-square error

$$\frac{\|A(X^k_d - X_0)\|_F}{\sqrt{1000209}},$$

which does not decrease to zero, but represents how closely the current iterate fits the observed data.

The code runs fairly quick for the scale of the data. The main bottleneck in this algorithm is evaluating the proximal operator of $\| \cdot \|_*$, which requires computing the SVD of a $6040 \times 3952$ size matrix.

6.5.3 Support vector machine classification

In support vector machine classification we have a kernel matrix $K \in \mathbb{R}^{d \times d}$ generated from a training set $X = \{t_1, \ldots, t_d\}$ using a kernel function $\mathcal{K} : X \times X \rightarrow \mathbb{R}$: for all $i, j = 1, \ldots, d$, we have $K_{i,j} = \mathcal{K}(t_i, t_j)$. In our particular example, $X \subseteq \mathbb{R}^n$ for some $n > 0$ and $\mathcal{K}_\sigma : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_{++}$ is the Gaussian kernel given by $\mathcal{K}_\sigma(t, t') = e^{-\sigma \|t - t'\|^2}$ for some $\sigma > 0$. We are also given a label vector $y \in \{-1, 1\}^d$, which indicates the label given to each point in $X$. Finally, we are given a real number $C > 0$ that controls how much we let our final classifier stray from perfect classification on the training set $X$.

We define constraint sets $\mathcal{C}_1 = \{0 \leq x \leq C\}$ and $\mathcal{C}_2 = \{x \in \mathbb{R}^d \mid \langle y, x \rangle = 0\}$. We also define $Q_0 = \text{diag}(y)K\text{diag}(y)$. Then the solution to Problem (6.2.10) with $Q = Q_0$
Figure 6.3: Run time and convergence rate statistics for the matrix completion Problem (6.2.9) on the MovieLens-1M database [1].
is precisely the dual form soft-margin SVM classifier [54]. Unfortunately, the Lipschitz constant of $Q_0$ is often quite large (i.e., $\gamma$ must be small), which results in poor practical performance. Thus, to improve practical we solve Problem (6.2.10) with $Q = P_{C_2}Q_0P_{C_2}$, which is equivalent to the original problem because the minimizer must lie in $C_2$. The result is a much smaller Lipschitz constant for $Q$ and better practical performance. This trick was first reported in Chapter 5.

We evaluated our algorithm on a subset $X_{\text{all}}$ of the UCI “Adult” machine learning dataset which is entitled “a7a” and is available from the LIBSVM website [37]. Our training set $X_{\text{train}}$ consisted of a $d = 9660$ element subsample of this 16100 element training set (i.e., a 60% sample). Note that $Q$ has $d^2 = 9660^2 = 93315600$ nonzero entries. In table 6.1, we trained the SVM model (6.2.10) with different choices of parameters $C$ and $\sigma$, and then evaluated their prediction accuracy on the remaining $16100 - 9660 = 6440$ elements in $X_{\text{test}} = X_{\text{all}} \setminus X_{\text{train}}$. We found that the parameters $\sigma = 2^{-3}$ and $C = 1$ gave the best performance on the test set, so we set these to be the parameters for our numerical experiments.

Figure 6.4 plots the results of our test. Figures 6.4a and 6.4b compare the line search method in Algorithm 6.1.3 with the basic Algorithm 6.1.1. We see that the line search method performs better than the basic algorithm in terms of number of iterations and total CPU time needed to reach a desired accuracy. Because of the linearity of the projection $P_{C_2}$, we can find a closed form solution for the line search weight $\rho$ in Algorithm 6.4a as the root of a third degree polynomial. Thus, although Algorithm 6.1.3 requires more work per iteration than Algorithm 6.1.1, it still takes less time overall because Algorithm 6.1.1 must compute $\beta = 1/\|Q\|$, which is quite costly.

Finally, in Figure 6.4c we compare the performance of the nonergodic iterate generated by Algorithm 6.1.1, the standard ergodic iterate (6.1.6), and the newly introduced ergodic iterate (6.1.7). We see that the nonergodic iterate performs better than the other two, and as expected, the new ergodic iterate outperforms the standard ergodic iterate. We emphasize that computing these iterates is essentially costless for the user and only
modifies the final output of the algorithm, not the trajectory.

We emphasize that all steps in this algorithm can be computed in closed form, so implementation is easy and each iteration is quite cheap.

6.5.4 Portfolio optimization

In this section, we evaluate our algorithm on the portfolio optimization problem. In this problem, we have a choice to invest in \( d > 0 \) assets and our goal is to choose how to distribute our resources among all the assets so that we minimize investment risk, and guarantee that our expected return on the investments is greater than \( r \geq 0 \). Mathematically, we model the distribution of our assets with a vector \( x \in \mathbb{R}^d \) where \( x_i \) represents the percentage of our resources that we invest in asset \( i \). For this reason, we define our constraint set \( C_1 = \{ x \in \mathbb{R}^d \mid \sum_{i=1}^n x_i = 1, x_i \geq 0 \} \) to be the standard simplex. We also assume that we are given a vector of mean returns \( m \in \mathbb{R}^d \) where \( m_i \) represents the expected return from asset \( i \), and we define \( C_2 = \{ x \in \mathbb{R}^d \mid \langle m, x \rangle \geq r \} \). Typically, we model the risk with a matrix \( Q_0 \in \mathbb{R}^{d \times d} \), which is usually chosen as the covariance matrix.

<table>
<thead>
<tr>
<th>( C )</th>
<th>kernel parameter ( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 2^{-5} )</td>
<td>0.82689</td>
</tr>
<tr>
<td>( 2^{-3} )</td>
<td>0.82658</td>
</tr>
<tr>
<td>( 2^{-1} )</td>
<td>0.83465</td>
</tr>
<tr>
<td>( 2 )</td>
<td>0.83465</td>
</tr>
</tbody>
</table>

Table 6.1: Classification accuracy for different choices of \( C \) and \( \sigma \) in the SVM model.
Figure 6.4: Run time and convergence rate statistics for the SVM Problem (6.2.10) on the UCI “Adult” Machine learning dataset [83]. Results are with the parameter choice that has the best generalization to the test set $(C, \sigma) = (1, 0.2^{-3})$. 

(a) Fixed-point residual with and without line search (LS).

(b) Objective value with and without line search (LS).

(c) Comparison of ergodic and nonergodic iterates.
of asset returns. However, we stray from the typical model by setting $Q = Q_0 + \mu I_{R^d}$ for some $\mu \geq 0$, which has the effect of encouraging diversity of investments among the assets. In order to choose our optimal investment strategy, we solve Problem (6.2.10) with $Q$, $C_1$ and $C_2$ introduced here.

In our numerical experiments, we solve a $d = 1000$ dimensional portfolio optimization problem with a randomly generated covariance matrix $Q_0$ (using the Matlab “gallery” function) and mean return vector $m$. We report our results in Figure 6.5. In order to get an estimate of the solution of Problem (6.2.10), we first solved this problem to high-accuracy using an interior point solver.

The matrix $Q$ in this example is positive definite for any choice of $\mu \geq 0$, but the condition number of $Q_0$ is around 8000, while the condition number of $Q$ with $\mu = .1$ is around 5. For this reason, we see a huge improvement in Figure 6.5a with the acceleration in Algorithm 6.1.2, while in the case $\mu = 0$ in Figure 6.5b, the accelerated and non accelerated versions are nearly identical.

We emphasize that all steps in this algorithm can be computed in nearly closed form, so implementation is easy and each iteration is quite cheap.

6.6 Conclusion

In this chapter, we introduced a new operator-splitting algorithm for the three-operator monotone inclusion problem, which has a large variety of applications. We showed how to accelerate the algorithm whenever one of the involved operators is strongly monotone, and we also introduced a line search procedure and two averaging strategies that can improve the convergence rate. We characterized the convergence rate of the algorithm under various scenarios and showed that many of our rates are sharp. Finally, we introduced numerous applications of the algorithm and showed how it unifies many existing splitting schemes.
Appendices

A Proofs of technical results from Chapter 2

A.1 Proofs from Section 2.8

**Proposition 2.8.1.** By Equation (2.8.3) and Lemma 2.8.1, we get the following formulation for the $k$-th iteration: Given $z^0 \in \mathcal{H}$

\[
\begin{align*}
    y^k &= \arg \min_{y \in \mathcal{H}_2} g(y) - \langle z^k, By - b \rangle + \gamma \frac{1}{2} \|By - b\|^2 \\
    w_{d_g}^k &= z^k - \gamma(By^k - b) \\
    x^k &= \arg \min_{x \in \mathcal{H}_1} f(x) - \langle 2w_{d_g}^k - z^k, Ax \rangle + \gamma \frac{1}{2} \|Ax\|^2 \\
    w_{d_f}^k &= 2w_{d_g}^k - z^k - \gamma Ax^k \\
    z^{k+1} &= z^k + 2\lambda_k(w_{d_f}^k - w_{d_g}^k)
\end{align*}
\]

We will use this form to get to the claimed iteration. First,

\[
2w_{d_g}^k - z^k = w_{d_g}^k - \gamma(By^k - b) \quad \text{and} \quad w_{d_f}^k = w_{d_g}^k - \gamma(Ax^k + By^k - b). \tag{A.2}
\]
Furthermore, we can simplify the definition of $x^k$:

$$x^k = \arg \min_{x \in H_1} \left\{ f(x) - \langle 2w_{dy}^k - z^k, Ax \rangle + \frac{\gamma}{2} \|Ax\|^2 \right\}$$

(E.2)\[= \arg \min_{x \in H_1} \left\{ f(x) - \langle w_{dy}^k - \gamma(By^k - b), Ax \rangle + \frac{\gamma}{2} \|Ax\|^2 \right\}
= \arg \min_{x \in H_1} \left\{ f(x) - \langle w_{dy}^k, Ax + By^k - b \rangle + \frac{\gamma}{2} \|Ax + By^k - b\|^2 \right\}. \quad (A.3)$$

Note that the last two lines of Equation (A.3) differ by terms independent of $x$.

We now eliminate the $z^k$ variable from the $y^k$ subproblem: because $w_{df}^k + z^k = 2w_{dy}^k - \gamma Ax^k$, we have

$$z^{k+1} = z^k + 2\lambda_k (w_{df}^k - w_{dy}^k)$$

(A.2)\[= z^k + w_{df}^k - w_{dy}^k + \gamma(2\lambda_k - 1)(Ax^k + By^k - b)
= w_{dy}^k - \gamma Ax^k - \gamma(2\lambda_k - 1)(Ax^k + By^k - b). \quad (A.4)$$

We can simplify the definition of $y^{k+1}$ by with the identity in Equation (A.4):

$$y^{k+1} = \arg \min_{y \in H_2} \left\{ g(y) - \langle z^{k+1}, By - b \rangle + \frac{\gamma}{2} \|By - b\|^2 \right\}
= \arg \min_{y \in H_2} \left\{ g(y) - \langle w_{dy}^k - \gamma Ax^k - \gamma(2\lambda_k - 1)(Ax^k + By^k - b), By - b \rangle
+ \frac{\gamma}{2} \|By - b\|^2 \right\}
= \arg \min_{y \in H_2} \left\{ g(y) - \langle w_{dy}^k, Ax^k + By - b \rangle
+ \frac{\gamma}{2} \|Ax^k + By - b + (2\lambda_k - 1)(Ax^k + By^k - b)\|^2 \right\}. \quad (A.5)$$

The result then follows from Equations (A.1), (A.2), (A.3), and (A.5), combined with the initial conditions listed in the statement of the proposition. In particular, note that the updates of $x, y, w_{df},$ and $w_{dy}$ do not explicitly depend on $z$.

The following proposition will help us derive primal fundamental inequalities akin to Proposition 2.4.1 and 2.4.2.
Proposition A.1. Suppose that \((z^j)_{j \geq 0}\) is generated by Algorithm 2. Let \(z^*\) be a fixed point of \(T_{PRS}\) and let \(w^* = \text{prox}_{\gamma_d,f}(z^*)\). Then the following identity holds:

\[
4\gamma \lambda_k (f(x^k) + g(y^k) - f(x^*) - g(y^*)) = -4\gamma \lambda_k (d_f(w^k_{d_f}) + d_g(w^k_{d_g}) - d_f(w^*) - d_g(w^*)) + \left( 2 \left( 1 - \frac{1}{2\lambda_k} \right) \|z^k - z^{k+1}\|^2 + 2\langle z^k - z^{k+1}, z^{k+1}\rangle \right). \tag{A.6}
\]

Proof. We have the following subgradient inclusions from Proposition 2.8.1: \(A^*w^k_{d_f} \in \partial f(x^k)\) and \(B^*w^k_{d_g} \in \partial g(y^k)\). From the Fenchel-Young inequality \([11, \text{Proposition 16.9}]\) we have the expression for \(f\) and \(g\):

\[
d_f(w^k_{d_f}) = \langle A^*w^k_{d_f}, x^k \rangle - f(x^k) \quad \text{and} \quad d_f(w^k_{d_f}) = \langle B^*w^k_{d_g}, y^k \rangle - g(y^k) - \langle w^k_{d_g}, b \rangle.
\]

Therefore,

\[
-d_f(w^k_{d_f}) - d_g(w^k_{d_g}) = f(x^k) + g(y^k) - \langle Ax^k + By^k - b, w^k_{d_f} \rangle - \langle w^k_{d_g} - w^k_{d_f}, By^k - b \rangle.
\]

Let us simplify this bound with an identity from Proposition 2.8.1: from \(w^k_{d_f} - w^k_{d_g} = -\gamma(Ax^k + By^k - b)\), it follows that

\[
-d_f(w^k_{d_f}) - d_g(w^k_{d_g}) = f(x^k) + g(y^k) + \frac{1}{\gamma} \langle w^k_{d_f} - w^k_{d_g}, w^k_{d_f} + \gamma(By^k - b) \rangle. \tag{A.7}
\]

Recall that \(\gamma(By^k - b) = z^k - w^k_{d_g}\). Therefore

\[
w^k_{d_f} + \gamma(By^k - b) = z^k + (w^k_{d_f} - w^k_{d_g}) = z^k + \frac{1}{2\lambda_k}(z^{k+1} - z^k) = \frac{1}{2\lambda_k}(2\lambda_k - 1)(z^k - z^{k+1}) + z^{k+1},
\]

and the inner product term can be simplified as follows:

\[
\frac{1}{\gamma} \langle w^k_{d_f} - w^k_{d_g}, w^k_{d_f} + \gamma(By^k - b) \rangle = \frac{1}{\gamma} \left( \frac{1}{2\lambda_k}(z^{k+1} - z^k), \frac{1}{2\lambda_k}(2\lambda_k - 1)(z^k - z^{k+1}) \right) + \frac{1}{\gamma} \left( \frac{1}{2\lambda_k}(z^{k+1} - z^k), z^{k+1} \right)
\]

\[
= -\frac{1}{2\gamma\lambda_k} \left( 1 - \frac{1}{2\lambda_k} \right) \|z^{k+1} - z^k\|^2
\]

\[
= -\frac{1}{2\gamma\lambda_k} \langle z^k - z^{k+1}, z^{k+1} \rangle. \tag{A.8}
\]
Now we derive an expression for the dual objective at a dual optimal \( w^* \). First, if \( z^* \) is a fixed point of \( T_{\text{PRS}} \), then 
\[
0 = T_{\text{PRS}}(z^*) - z^* = 2(w^*_d - w^*_d) = -2\gamma(Ax^* + By^* - b).
\]
Thus, from Equation (A.7) with \( k \) replaced by \( * \), we get 
\[
-d_f(w^*) - d_g(w^*) = f(x^*) + g(y^*) + \langle Ax^* + Bx^* - b, w^* \rangle = f(x^*) + g(y^*). \tag{A.9}
\]

Therefore, Equation (A.6) follows by subtracting (A.9) from Equation (A.7), rearranging and using the identity in Equation (A.8).

Proposition 2.8.2. The fundamental lower inequality in Proposition 2.4.2 applied to \( d_f + d_g \) shows that 
\[
-4\gamma\lambda_k(d_f(w^*_d) + d_g(w^*_d)) - d_f(w^*) - d_g(w^*) \leq 2\langle z^{k+1} - z^k, z^* - w^* \rangle.
\]
The proof then follows from Proposition A.1, and the simplification:
\[
2\langle z^k - z^{k+1}, z^{k+1} - (z^* - w^*) \rangle + 2\left(1 - \frac{1}{2\lambda_k}\right)\|z^k - z^{k+1}\|^2 \\
= \|z^k - (z^* - w^*)\|^2 - \|z^{k+1} - (z^* - w^*)\|^2 + \left(1 - \frac{1}{\lambda_k}\right)\|z^k - z^{k+1}\|^2.
\]
B Proofs of technical results from Chapter 5

B.1 Proof of Theorem 5.1.2

For the proof, we ask the reader to recall (5.1.16).

For all \( k \geq 0 \), set

\[
p^k = \frac{1 - \alpha_1}{\alpha_1} \|(I_H - T_1) \circ T_2(z^k) - (I_H - T_1) \circ T_2(z^*)\|^2 + \frac{1 - \alpha_2}{\alpha_2} \|(I_H - T_2)(z^k) - (I_H - T_2)(z^*)\|^2.
\]

By applying (5.1.14) twice, we get \( \|T_1 \circ T_2(z^k) - T_1 \circ T_2(z^*)\|^2 \leq \|z^k - z^*\|^2 - p^k \).

Part 5 of Proposition 5.1.1 shows that \( (T_1 \circ T_2)_{\lambda_k} \) is \( (\alpha_{1,2}\lambda_k) \)-averaged. Thus,

\[
\|z^{k+1} - z^*\|^2 \leq \|z^k - z^*\|^2 - \frac{\lambda_k(1 - \lambda_k\alpha_{1,2})}{\alpha_{1,2}} \|T_1 \circ T_2(z^k) - z^k\|^2.
\]

Therefore, \( \sum_{i=0}^{\infty} \lambda_i(1 - \alpha_{1,2}\lambda_i) \|T_1 \circ T_2(z^i) - z^i\|^2 \leq \|z^0 - z^*\|^2 \).

By [11, Corollary 2.14], the following holds: for all \( x, y \in \mathcal{H} \) and all \( \lambda \in \mathbb{R} \), we have

\[
\|\lambda x + (1 - \lambda)y\|^2 = \lambda\|x\|^2 + (1 - \lambda)\|y\|^2 - \lambda(1 - \lambda)\|x - y\|^2.
\]

Therefore, we have

\[
\|z^{k+1} - z^*\|^2 = (1 - \lambda_k)\|z^k - z^*\|^2 + \lambda_k\|T_1 \circ T_2(z^k) - T_1 \circ T_2(z^*)\|^2 - \lambda_k(1 - \lambda_k)\|z^k - T_1 \circ T_2(z^*)\|^2
\]
\[
\leq \|z^k - z^*\|^2 - \lambda_k p^k + \lambda_k(\lambda_k - 1)\|z^k - T_1 \circ T_2(z^*)\|^2
\]
\[
\leq \|z^k - z^*\|^2 - \lambda_k p^k + \frac{\lambda_k(1 - \alpha_{1,2}\lambda_k)}{\alpha_{1,2}\varepsilon} \|z^k - T_1 \circ T_2(z^*)\|^2.
\]

Thus, take \( k \to \infty \) in the following inequality to get the result:

\[
\sum_{i=0}^{k} \lambda_i\|T_1 \circ T_2(z^i) - (I_H - T_2)(z^*)\|^2 \leq \frac{\alpha_2}{1 - \alpha_2} \sum_{i=0}^{k} \lambda_i p^i
\]
\[
\leq \frac{\alpha_2}{1 - \alpha_2} \sum_{i=0}^{k} \left( \|z^i - z^*\|^2 - \|z^{i+1} - z^*\|^2 + \frac{\lambda_i(1 - \alpha_{1,2}\lambda_i)}{\alpha_{1,2}\varepsilon} \|z^i - T_1 \circ T_2(z^i)\|^2 \right)
\]
\[
\leq \alpha_2(1 + 1/\varepsilon) \|z^0 - z^*\|^2.
\]
B.2 Proof of Lemma 5.2.1

The identity for \( x_h = z - \gamma \tilde{\nabla} \chi_V(x_h) \) follows from Part 1 of Proposition 5.1.1. Note that by the Moreau identity \( P_{V^\perp} = I - P_V \), we have \( \gamma \tilde{\nabla} \chi_V(x_h) = P_{V^\perp} z \). Note that by definition, \( \nabla h(z) = P_V \circ \nabla g \circ P_V(z) = P_V \circ \nabla g(x_h) = \nabla h(x_h) \) and \( \nabla h(z) \in V \). Thus, we get the identity for \( x_f \):

\[
\text{prox}_{\gamma f} \circ \text{refl}_{\chi_V} \circ (I_{\mathcal{H}} - \gamma \nabla h)(z) = \text{refl}_{\chi_V} \circ (I_{\mathcal{H}} - \gamma \nabla h)(z) - \gamma \tilde{\nabla} f(x_f)
\]

\[
= x_h - \gamma \nabla h(z) - P_{V^\perp} z - \gamma \tilde{\nabla} f(x_f) = x_h - \gamma \left( \tilde{\nabla} \chi_V(x_h) + \nabla h(x_h) + \tilde{\nabla} f(x_f) \right).
\]

Finally, given the identity \( (T_{\text{FDRS}})_\lambda(z) - z = \lambda(T_{\text{FDRS}}(z) - z) \), (5.2.2) will follow as soon as we show \( T_{\text{FDRS}}(z) = x_f + z - x_h = x_f + \gamma \tilde{\nabla} \chi_V(x_h) \):

\[
\left( \frac{1}{2} I_{\mathcal{H}} + \frac{1}{2} \text{refl}_{\gamma f} \circ \text{refl}_{\chi_V} \right) (z - \gamma \nabla h(z)) = (\text{prox}_{\gamma f} \circ \text{refl}_{\chi_V} + I_{\mathcal{H}} - P_V) (z - \gamma \nabla h(z))
\]

\[
= x_f + P_{V^\perp} (z - \gamma \nabla h(z)) = x_f + \gamma \tilde{\nabla} \chi_V(x_h).
\]

B.3 Proof of Lemma 5.2.2

Let \( x \in \text{zer}(\partial f + \nabla h + \partial \chi_V) \). Choose subgradients \( \tilde{\nabla} f(x) \in \partial f(x) \) and \( \tilde{\nabla} \chi_V(x) \in \partial \chi_V(x) = V^\perp \) (by (5.1.9)) such that \( \tilde{\nabla} f(x) + \nabla h(x) + \tilde{\nabla} \chi_V(x) = 0 \) and set \( z := x + \gamma \tilde{\nabla} \chi_V(x) \). We claim that \( z \) is a fixed-point of \( T_{\text{FDRS}} \). From Lemma 5.2.1, we get the points:

\( x_h := P_V(z) = x \) and \( x_f := \text{prox}_{\gamma f} \circ \text{refl}_{\chi_V} \circ (I_{\mathcal{H}} - \gamma \nabla h)(z) \). But \( \tilde{\nabla} \chi_V(x_h) + \nabla h(x_h) \in -\partial f(x) \), and

\[
\text{refl}_{\chi_V} \circ (I_{\mathcal{H}} - \gamma \nabla h)(z) = P_V(z - \gamma \nabla h(z)) + (P_V - I_{\mathcal{H}})(z - \gamma \nabla h(z))
\]

\[
= x - \gamma \nabla h(x) - P_{V^\perp} z = x - \gamma \nabla h(x) - \gamma \tilde{\nabla} \chi_V(x) = x + \gamma \tilde{\nabla} f(x).
\]

Therefore, \( x_f = \text{prox}_{\gamma f}(x + \gamma \tilde{\nabla} f(x)) = x = x_h \) (see Part 1 of Proposition 5.1.1). Thus, by Lemma 5.2.1, \( T_{\text{FDRS}} z = z + x_f - x_h = z \). We have proved the first inclusion.

On the other hand, suppose that \( z \in \mathcal{H} \) and \( T_{\text{FDRS}} z = z \). Then \( x := x_h = P_V z \), and

\[
0 = T_{\text{FDRS}} z - z = x_f - x_h = -\gamma \left( \tilde{\nabla} \chi_V(x_h) + \nabla h(x_h) + \tilde{\nabla} f(x_f) \right).
\]

Because \( x_f = x_h \), we get \( x \in \text{zer}(\partial f + \nabla h + \partial \chi_V) \).

258
B.4 Proof of Proposition 5.2.1

In the following derivation, we use (5.2.3) and (5.2.4), Lemma 5.2.1, the cosine rule, and the inclusion $\tilde{\nabla} \chi_V(x_h) \in V^\perp$:

$$2\gamma \lambda (f(x_f) + h(x_h) - f(x) - h(x) + S_f(x_f, x) + S_h(x_h, x))$$

$$\leq 2\gamma \lambda \left( \langle \tilde{\nabla} f(x_f), x_f - x \rangle + \langle \nabla h(x_h), x_h - x \rangle + \langle \tilde{\nabla} \chi_V(x_h), x_h - x \rangle \right)$$

$$= 2\gamma \lambda \left( \langle \tilde{\nabla} f(x_f) + \nabla h(x_h) + \tilde{\nabla} \chi_V(x_h), x_f - x \rangle + \langle \nabla h(x_h) + \tilde{\nabla} \chi_V(x_h), x_h - x \rangle \right)$$

$$= 2\langle z - z^+, x_f - x \rangle + 2\langle \gamma \nabla h(x_h) + \gamma \tilde{\nabla} \chi_V(x_h), z - z^+ \rangle$$

$$= 2\langle z - z^+, x_f + \gamma \tilde{\nabla} \chi_V(x_h) - x \rangle + 2\gamma \langle \nabla h(x_h), z - z^+ \rangle$$

$$= 2\langle z - z^+, T_{FDRS} z - x \rangle + 2\gamma \langle \nabla h(x_h), z - z^+ \rangle$$

$$= 2\langle z - z^+, z - x \rangle + \frac{2}{\lambda} \langle z - z^+, z^+ - z \rangle + 2\gamma \langle \nabla h(x_h), z - z^+ \rangle$$

$$= \langle z - x \rangle^2 - \langle z^+ - x \rangle^2 + \left( 1 - \frac{2}{\lambda} \right) \langle z^+ - z \rangle^2 + 2\gamma \langle \nabla h(x_h), z - z^+ \rangle. \tag{5.1.11}$$

B.5 Proof of Proposition 5.2.2

By the subgradient inequality and because $\tilde{\nabla} \chi_V(x_h) \in V^\perp$, we have

$$f(x_f) + h(x_h) - f(x^*) - g(x^*) \leq \langle x_h - x^*, \tilde{\nabla} f(x^*) + \nabla h(x^*) + \tilde{\nabla} \chi_V(x^*) \rangle$$

$$+ \langle x_f - x_h, \tilde{\nabla} f(x^*) \rangle + S_f(x_f, x^*) + S_h(x_h, x^*)$$

$$= \langle x_f - x_h, \tilde{\nabla} f(x^*) \rangle + S_f(x_f, x^*) + S_h(x_h, x^*).$$

B.6 Proof of Corollary 5.2.1

By (5.1.11), we have $\|z - x^*\|^2 - \|z^* - x^*\|^2 = \|z - z^*\|^2 - \|z^+ - z^*\|^2 + 2\langle z - z^+, z^* - x^* \rangle$.

Therefore, by Proposition 5.2.1,

$$2\gamma \lambda (f(x_f) + h(x_h) - f(x^*) - h(x^*) + S_f(x_f, x^*) + S_h(x_h, x^*))$$

$$\leq \|z - z^*\|^2 - \|z^+ - z^*\|^2 + 2\langle z - z^+, z^* - x^* \rangle$$

$$+ \left( 1 - \frac{2}{\lambda} \right) \|z^+ - z\|^2 + 2\gamma \langle \nabla h(x_h), z - z^+ \rangle. \tag{B.1}$$
Equation (5.2.9) now follows from (B.1) and (5.2.8):

\[ 4\gamma \lambda (S_f(x_f, x^*) + S_h(x_h, x^*)) \leq -2\gamma \lambda_k \langle x_f - x_h, \nabla f(x^*) \rangle + 2\gamma \lambda_k (f(x_f) + h(x_h) - f(x^*) - h(x^*) + S_f(x_f, x^*) + S_h(x_h, x^*)) \]

\[ \leq \|z - z^*\|^2 - \|z^+ - z^*\|^2 + 2\gamma \lambda_k (\langle x_f - x_h, \nabla f(x^*) \rangle + \left(1 - \frac{2}{\lambda}\right) \|z^+ - z\|^2 + 2\gamma \langle \nabla h(x_h), z - z^+ \rangle) \]

\[ \equiv \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left(1 - \frac{2}{\lambda}\right) \|z^+ - z\|^2 + 2\gamma \langle \nabla h(x_h), \nabla h(x^*), z - z^+ \rangle. \]

### B.7 Proof of Theorem 5.4.1

Let \( \eta_k = 2/\lambda_k - 1 \). By (5.3.3), we have

\[ 2\gamma \langle \nabla h(x_h^k) - \nabla h(x^*), z^k - z^{k+1} \rangle \leq \frac{\gamma^2}{\eta_k} \|\nabla h(x_h^k) - \nabla h(x^*)\|^2 + \eta_k \|z^k - z^{k+1}\|^2. \] (B.2)

Hence, for all \( k \geq 0 \), we have (using \( 1/\eta_k \leq \lambda_k/\varepsilon^2 \) as in (5.3.3) and (5.1.16))

\[ 4\gamma \lambda \sum_{i=0}^{k} (S_f(x_f^i, x^*) + S_h(x_h^i, x^*)) \leq \sum_{i=0}^{k} 4\gamma \lambda_i (S_f(x_f^i, x^*) + S_h(x_h^i, x^*)) \]

\[ \leq \sum_{i=0}^{k} \left( \|z^i - z^*\|^2 - \|z^{i+1} - z^*\|^2 - \eta_i \|z^{i+1} - z^i\|^2 \right. \]

\[ + 2\gamma \langle \nabla h(x_h^i), \nabla h(x^*), z^i - z^{i+1} \rangle \]

\[ \leq \sum_{i=0}^{k} \left( \|z^i - z^*\|^2 - \|z^{i+1} - z^*\|^2 + (\gamma^2 \lambda_i/\varepsilon^2) \|\nabla h(x_h^i) - \nabla h(x^*)\|^2 \right) \]

\[ \leq \|z^0 - z^*\|^2 - \|z^{k+1} - z^*\|^2 + \frac{(1 + \varepsilon)\gamma^3}{\varepsilon^3(2\beta_V - \gamma)} \|z^0 - z^*\|^2. \]

The “best” convergence rates now follow by taking \( k \to \infty \) and using Lemma 2.2.1 of Chapter 2. In addition, we apply Jensen’s inequality to the convex function \( \|\cdot\|^2 \) to get

\[ \overline{S}_f^k + \overline{S}_h^k \leq \frac{1}{\Lambda_k} \sum_{i=0}^{k} \lambda_i (S_f(x_f^i, x^*) + S_h(x_h^i, x^*)) \leq \left(1 + \frac{(1 + \varepsilon)\gamma^3}{\varepsilon^3(2\beta_V - \gamma)}\right) \frac{\|z^0 - z^*\|^2}{4\gamma \Lambda_k}. \]

Now, for all \( \lambda > 0 \), define \( z_\lambda = (T_{\text{FDRS}})_\lambda(z^k) \). Observe that \( S_f(x_f^k, x^*) \) and \( S_h(x_h^k, x^*) \)
do not depend on the value of $\lambda_k$. Therefore, we use (5.2.9) to get

\[
S_f(x_f^k, x^*) + S_h(x_h^k, x^*) \leq \inf_{\lambda \in [0, 1/\alpha_{\text{FDRS}}]} \frac{1}{2\gamma \lambda} \left( 2\gamma \langle \nabla h(x_h^k) - \nabla h(x^*), z^k - z_\lambda \rangle + \|z^k - z^*\|^2 - \|z_\lambda - z^*\|^2 + \left( 1 - \frac{2}{\lambda} \right) \|z_\lambda - z^k\|^2 \right)
\]

(5.1.11)

\[
= \inf_{\lambda \in [0, 1/\alpha_{\text{FDRS}}]} \frac{1}{2\gamma \lambda} \left( 2\gamma \langle \nabla h(x_h^k) - \nabla h(x^*), z^k - z_\lambda \rangle + 2\langle z_\lambda - z^*, z^k - z_\lambda \rangle + 2 \left( 1 - \frac{1}{\lambda} \right) \|z_\lambda - z^k\|^2 \right)
\]

(5.1.19)

\[
\leq \frac{1}{2\gamma} \left( 2\langle z_1 - z^*, z^k - z_1 \rangle + \frac{2\gamma}{\beta_V} \|z^k - z^*\| \|z_1 - z_k\| \right)
\]

(B.3)

where the (B.3) uses the $1/\beta_V$-Lipschitz continuity of $\nabla h$ and the identity $\nabla h(x_h^k) - \nabla h(x^*) = \nabla h(z^k) - \nabla h(z^*)$, and the last line uses the Fejér property $\|z_1 - z^*\| \leq \|z^k - z^*\| \leq \|z^0 - z^*\|$ (see Part 1 of Theorem 5.1.1). The $o(1/\sqrt{k+1})$ rates follow from (B.3) and the corresponding rates for the FPR in (5.1.19).

### B.8 Proof of Proposition 5.5.1

Because $\nabla f$ is $(1/\beta_f)$-Lipschitz, we have

\[
f(x_h) \leq f(x_f) + \langle x_h - x_f, \nabla f(x_f) \rangle + \frac{1}{2\beta_f} \|x_h - x_f\|^2; \quad (B.4)
\]

\[
S_f(x_f, x^*) \geq \frac{\beta_f}{2} \|\nabla f(x_f) - \nabla f(x^*)\|^2. \quad (B.5)
\]

where the first inequality follows from [11, Theorem 18.15(iii)]. By applying the identity $z^* - x^* = \gamma \widetilde{\nabla}_V(x^*) = -\gamma \nabla f(x^*) - \gamma \nabla h(x^*)$, the cosine rule (5.1.11), and the identity
\[ z - z^+ = \lambda(x_h - x_f) \text{ (see (5.2.2)) multiple times, we have} \]

\[
2(z - z^+, z^* - x^*) + 2\gamma\lambda(x_h - x_f, \nabla f(x_f)) = 2\lambda(x_h - x_f, \gamma \nabla \chi_V(x^*) + \gamma \nabla f(x_f))
\]

\[
= 2\lambda(\gamma \nabla \chi_V(x_h) + \gamma \nabla h(x_h) + \gamma \nabla f(x_f), \gamma \nabla f(x_f) - \gamma \nabla f(x^*)) - 2(z - z^+, \gamma \nabla h(x^*))
\]

\[
= \lambda \left( \left\| \gamma \nabla f(x_f) - \gamma \nabla f(x^*) \right\|^2 + \|x_h - x_f\|^2 
- \gamma^2 \| \nabla \chi_V(x_h) + \nabla h(x_h) - \nabla \chi_V(x^*) - \nabla h(x^*) \|^2 \right) - 2(z - z^+, \gamma \nabla h(x^*)). \tag{B.6}
\]

By (5.2.2) (i.e., \( z - z^+ = \lambda(x_h - x_f) \)), we have

\[
\left( 1 - \frac{2\lambda}{\beta_f} \right) \|z - z^+\|^2 + \lambda \left( \frac{\gamma}{\beta_f} + 1 \right) \|x_h - x_f\|^2 = \left( 1 + \frac{\gamma - \beta_f}{\beta_f \lambda} \right) \|z - z^+\|^2.
\]

Therefore,

\[
2\gamma\lambda(f(x_h) + h(x_h) - f(x^*) - h(x^*))
\]

\[
\leq 2\gamma\lambda(f(x_f) + h(x_h) - f(x^*) - h(x^*)) + 2\gamma\lambda(x_h - x_f, \nabla f(x_f)) + \frac{\gamma\lambda}{\beta_f} \|x_h - x_f\|^2 \tag{B.4}
\]

\[
\leq \|z - z^*\|^2 - \|z^+ - z^*\|^2 + 2(z - z^+, z^* - x^*) + 2\gamma\lambda(x_h - x_f, \nabla f(x_f))
+ \left( 1 - \frac{2\lambda}{\beta_f} \right) \|z - z^+\|^2 + \lambda \left( \frac{\gamma}{\beta_f} + 1 \right) \|x_h - x_f\|^2
\]

\[
\leq \left( 1 - \frac{2\lambda}{\beta_f} \right) \|z - z^+\|^2 + \gamma\lambda \|\nabla f(x_f) - \gamma \nabla f(x^*)\|^2 + 2\gamma \|\nabla h(x_h) - \nabla h(x^*), z - z^+\| - 2\gamma\lambda S_f(x_f, x^*)
\]

\[
\leq \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left( 1 + \frac{\gamma - \beta_f}{\beta_f \lambda} \right) \|z - z^+\|^2
+ 2\gamma \|\nabla h(x_h) - \nabla h(x^*), z - z^+\| + \gamma\lambda(\gamma - \beta_f) \|\nabla f(x_f) - \nabla f(x^*)\|^2. \tag{B.7}
\]

If \( \gamma \leq \beta_f \), then we can drop the last term. If \( \gamma > \beta_f \), then use (5.2.9) to get

\[
\gamma\lambda(\gamma - \beta_f) \|\nabla f(x_f) - \nabla f(x^*)\|^2 \leq \frac{(\gamma - \beta_f)}{2\beta_f} \left( 2\gamma \|\nabla h(x_h) - \nabla h(x^*), z - z^+\| + \|z - z^*\|^2 - \|z^+ - z^*\|^2 \right)
+ \left( 1 - \frac{2\lambda}{\beta_f} \right) \|z - z^+\|^2
\]

The result follows by (B.7) and

\[
\left( 1 + \frac{(\gamma - \beta_f)}{\beta_f \lambda} \right) \|z - z^+\|^2 + \frac{(\gamma - \beta_f)}{2\beta_f} \left( 1 - \frac{2\lambda}{\beta_f} \right) \|z - z^+\|^2 = \left( 1 + \frac{\gamma - \beta_f}{2\beta_f} \right) \|z - z^+\|^2.
\]
B.9 Proof of Theorem 5.6.3

For all $i \geq 0$, let $c_i = (i/(i + 1))^{1/2}$. Let $\kappa_a = (1/2) + 2(a + 1)^2$, and let $z^0 = \sqrt{2\alpha \kappa_a} e^{1/(a+1)} \times ((1/\|z_i\|)(i + 1)^{1/2})z_i \geq 0$. Then $\|z^0\|^2 = 2\alpha \kappa_a e^{2/(a+1)} \sum_{i=0}^{\infty} (1/(i + 1)^{2\alpha}) < \infty$ and, hence, $z^0 \in \mathcal{H}$. Now for all $i \geq 1$, we have

\[
\frac{\|z_i\|^2(a + c_i^2)^2}{c_i^2} \geq (1 - c_i^2) + \frac{(a + c_i^2)^2}{c_i^2} \leq \kappa_a
\]  

(B.8)

because $c_i^2 \in (1/2, 1)$. In addition, for all $i \geq 1$, we have

\[
\| (P_V)_i z^0_i \|^2 = \frac{2\alpha \kappa_a e^{2/(a+1)}}{\|z_i\|^2} \| (P_V)_i z^0_i \|^2 = \frac{2\alpha \kappa_a e^{2/(a+1)} c_i^2 (1 - c_i^2)}{\|z_i\|^2 (a + c_i^2)^2 (i + 1)^{2\alpha}} \\
= \frac{2\alpha \kappa_a e^{2/(a+1)} c_i^2}{\|z_i\|^2} \geq \frac{2\alpha e^{2/(a+1)}}{(i + 1)^{1+2\alpha}}
\]

where the third equality follows because $1 - c_i^2 = 1 - i/(i + 1) = 1/(i + 1)$.

Now, for all $k \geq 0$, let $z^{k+1} = T_{\text{FDRS}} z^k$. Again, for all $i \geq 0$, let $b_i = (a + c_i^2)/(a + 1)$ be the eigenvalue of $(T_{\text{FDRS}})_i$, associated to $z_i$. Note that $b_i^{2k} \geq e^{-2/(1+\alpha)}$ whenever $i \geq k \geq 0$.

Therefore, for all $k \geq 1$, we have

\[
\|x^k_h - x^*\|^2 = \|P_V T_{\text{FDRS}}^k z^0\|^2 = \sum_{i=0}^{\infty} b_i^{2(k+1)} \| (P_V)_i z^0_i \|^2 \geq \sum_{i=k}^{\infty} b_i^{2(k+1)} \frac{2\alpha e^{2/(a+1)}}{(i + 1)^{1+2\alpha}}
\]

where we use $x^* = 0$ and the lower integral approximation of the sum.

Now we prove the bound for $(x^k_f)_{j \geq 0}$. For all $k \geq 0$, $x^k_f = T_{\text{FDRS}} z^k - \gamma \nabla \chi_V(x^k_h) = T_{\text{FDRS}} z^k - P_{V^\perp} z^k = (T_{\text{FDRS}} - P_{V^\perp}) T_{\text{FDRS}} z^0$ (see (5.2.1)). In addition, for all $i \geq 0$,

\[
(T_{\text{FDRS}} - P_{V^\perp})_i = \frac{1}{(a + 1)} \begin{bmatrix}
0 & -\cos(\theta_j) \sin(\theta_j) \\
0 & \cos^2(\theta_j) + a - (a + 1) \end{bmatrix} = -\frac{\sin(\theta_j)}{(a + 1)} \begin{bmatrix}
0 & \cos(\theta_j) \\
0 & \sin(\theta_j) \end{bmatrix}.
\]

Thus, for all $i \geq 0$, we have

\[
\| (T_{\text{FDRS}} - P_{V^\perp})_i z^0_i \|^2 = \frac{2\alpha \kappa_a e^{2/(a+1)} \sin^2(\theta_j) (\cos^2(\theta_j) + \sin^2(\theta_j))}{\|z_i\|^2 (a + 1)^2 (i + 1)^{2\alpha}} = \frac{2\alpha \kappa_a e^{2/(a+1)} (1 - c_i^2)}{\|z_i\|^2 (a + 1)^2 (i + 1)^{2\alpha}} \geq \frac{2\alpha e^{2/(a+1)} (a + c_i^2)^2}{c_i^2 (a + 1)^2 (i + 1)^{1+2\alpha}}.
\]  

(B.8)
where the last inequality follows because $1 - c_i^2 = 1 - i/(i + 1) = 1/(i + 1)$ and $\kappa_a \| z_i \|^2 \geq (a + c_i^2)^2 / c_i^2$. Note that for all $i \geq 1$, we have $(a + c_i^2)^2 / c_i^2 \geq (a + 1/2)^2$ because $c_i^2 \in [1/2, 1)$.

Therefore, for all $k \geq 1$, we have

$$\| x_f^k - x^* \|^2 = \| (T_{\text{FDRS}} - P_{V^\perp})^k_{\text{FDRS}}^2 \|^2 \geq \sum_{i=k}^{\infty} b_i^{2(k+1)} \frac{2\alpha e^{2/(a+1)}(a + c_i^2)^2}{c_i^2(a + 1)^2(1 + i)^{1+2\alpha}} \geq \frac{(a + 1/2)^2}{(a + 1)^4(k + 1)^{2\alpha}}$$

where we use similar arguments to those used in (B.9).
C Proofs of technical results from Chapter 6

C.1 Proof of Theorem 6.1.2

We first prove a useful inequality.

**Proposition C.1.** Let $B$ be $\mu_B$-strongly monotone where we allow the case $\mu_B = 0$. Suppose that $x_0^A \in H$ and set $x_0^B = J_{\gamma_0 B}(x_0^A), u_0^B = (1/\gamma_0)(I - J_{\gamma_B})(x_0^A)$. For all $k \geq 0$, let

$$
\begin{align*}
  x^{k+1}_B &= J_{\gamma_k B}(x^k_A + \gamma_k u^k_B); \\
  u^{k+1}_B &= (1/\gamma_k)(x^k_A + \gamma_k u^k_B - x^{k+1}_B); \\
  x^{k+1}_A &= J_{\gamma_{k+1} A}(x^{k+1}_B - \gamma_{k+1} u^{k+1}_B - \gamma_{k+1} C x^{k+1}_B).
\end{align*}
$$

(C.1)

1. Suppose that $C$ is $\beta$-cocoercive and $\mu_C$-strongly monotone. Let $\eta \in (0,1)$ and let $(\gamma_j)_{j \geq 0} \subseteq (0, 2(1 - \eta)\beta)$. Then the following inequality holds for all $k \geq 0$:

$$
(1 + 2\gamma_k \mu_B)\|x^{k+1}_B - x^*\|^2 + \gamma_k^2 \|u^{k+1}_B - u^*_B\|^2 + \left(1 - \frac{\gamma_k}{2(1 - \eta)\beta}\right)\|x^k_A - x^*_B\|^2
\leq (1 - 2\gamma_k \mu_C \eta)\|x^k_B - x^*_B\|^2 + \gamma_k^2 \|u^k_B - u^*_B\|^2.
$$

(C.2)

2. Suppose that $C$ is $L_C$-Lipschitz, but not necessarily strongly monotone. In addition, suppose that $\mu_B > 0$. Then the following inequality holds for all $k \geq 0$:

$$
(1 + 2\gamma_k (\mu_B - \gamma_k L_C^2/2))\|x^{k+1}_B - x^*\|^2 + \gamma_k^2 L_C^2 \|x^{k+1}_B - x^*\|^2 + \gamma_k^2 \|u^{k+1}_B - u^*_B\|^2
\leq \|x^k_B - x^*_B\|^2 + \gamma_k^2 L_C^2 \|x^k_B - x^*_B\|^2 + \gamma_k^2 \|u^k_B - u^*_B\|^2.
$$

(C.3)

**Proof.** Fix $k \geq 0$.

Part 1: Following Fig. 6.1 and Lemma 6.3.1, let

$$
  u^k_A = \frac{1}{\gamma_k}(x^k_B - \gamma_k u^k_B - \gamma_k C x^k_B) - J_{\gamma_k A}(x^k_B - \gamma_k u^k_B - \gamma_k C x^k_B) \in Au^k_A.
$$
In addition, $u_B^k \in B u_B^k$ for all $k \geq 0$. The following identities from Fig. 6.1 will be useful in the proof:

$$
x_A^k - x_B^{k+1} = \gamma_k (u_B^{k+1} - u_B^k)
$$

$$
x_B^k - x_B^{k+1} = \gamma_k (u_B^{k+1} + C x_B^k + u_A^k)
$$

$$
x_B^k - x_A^k = \gamma_k (u_B^k + C x_B^k + u_A^k).
$$

First we bound the sum of two inner product terms.

$$
2\gamma_k \left( \langle x_A^k - x^*, u_A^k + C x_B^k \rangle + \langle x_B^{k+1} - x^*, u_B^{k+1} \rangle \right)
= 2\gamma_k \left( \langle x_A^k - x_B^{k+1}, u_A^k + C x_B^k \rangle + \langle x_B^{k+1} - x^*, u_B^{k+1} + u_A^k + C x_B^k \rangle \right)
= 2\gamma_k \left( \langle x_A^k - x_B^{k+1}, u_A^k + C x_B^k + u_B^k \rangle + \langle x_A^k - x_B^{k+1}, u_B^k \rangle \right)
+ 2\gamma_k \langle x_B^{k+1} - x^*, u_A^k + C x_B^k + u_B^k \rangle
= \|x_B^k - x_B^{k+1}\|^2 - \|x_A^k - x_B^{k+1}\|^2 - \|x_A^k - x_B^k\|^2
+ \|x_B^k - x_B^k\|^2 - \|x_B^{k+1} - x^*\|^2 - \|x_B^k - x_B^{k+1}\|^2
+ 2\gamma_k^2 \langle u_B^k - u_B^{k+1}, u_B^k - u_B^k \rangle + 2\gamma_k \langle x_A^k - x_B^{k+1}, u_B^k \rangle
= \|x_B^k - x_B^k\|^2 - \|x_B^{k+1} - x^*\|^2 - \|x_A^k - x_B^{k+1}\|^2 - \|x_A^k - x_B^k\|^2
+ \gamma_k^2 \left( \|u_B^k - u_B^k\|^2 - \|u_B^{k+1} - u_B^k\|^2 + \|u_B^k - u_B^{k+1}\|^2 \right) + 2\gamma_k \langle x_B^{k+1} - x_A^k, u_B^k \rangle
= \|x_B^k - x_B^k\|^2 - \|x_B^{k+1} - x^*\|^2 - \|x_A^k - x_B^k\|^2
+ \gamma_k^2 \|u_B^k - u_B^k\|^2 - \gamma_k^2 \|u_B^{k+1} - u_B^k\|^2 + 2\gamma_k \langle x_B^{k+1} - x_A^k, u_B^k \rangle. \quad (C.4)
$$

Furthermore, we have the lower bound

$$
2\gamma_k \left( \langle x_A^k - x^*, u_A^k + C x_B^k \rangle + \langle x_B^{k+1} - x^*, u_B^{k+1} \rangle \right)
\geq 2\gamma_k \left( \langle x_A^k - x^*, u_A^k + C x_B^k \rangle + \langle x_B^{k+1} - x^*, u_B^k \rangle \right) + 2\gamma_k \mu_B \|x_B^{k+1} - x^*\|^2. \quad (C.5)
$$

266
We have the further lower bound: For all \( \eta \in (0, 1) \), we have

\[
2\langle x_A^k - x^*, Cx_B^k \rangle = 2\langle x_A^k - x^*_B, Cx_B^k - Cx^* \rangle + 2\langle x_A^k - x^*_B, Cx^* \rangle + 2\langle x_B^k - x^*, Cx_B^k \rangle \\
\geq -\frac{1}{2\beta(1 - \eta)}\|x_A^k - x^*_B\|^2 - 2\beta(1 - \eta)\|Cx_B^k - Cx^*\|^2 \\
+ 2\mu_C\eta\|x_B^k - x^*\|^2 + 2\beta(1 - \eta)\|Cx_B^k - Cx^*\|^2 + 2\langle x_A^k - x^*, Cx^* \rangle.
\]

Altogether, we have

\[
2\gamma_k \left( \langle x_A^k - x^*, u_A^k + Cx_B^k \rangle + \langle x_B^{k+1} - x^*, u_B^{k+1} \rangle \right) \\
\geq 2\gamma_k \left( \langle x_A^k - x^*, u_A^k + Cx^* \rangle + \langle x_B^{k+1} - x^*, u_B^* \rangle \right) \\
- \frac{\gamma_k}{2(1 - \eta)\beta}\|x_A^k - x^*_B\|^2 + 2\gamma_k\mu_C(1 - \eta)\|x_B^k - x^*\|^2 + 2\gamma_k\mu_B\|x_B^{k+1} - x^*\|^2 \\
= 2\gamma_k\langle x_B^{k+1} - x_A^k, u_B^* \rangle - \frac{\gamma_k}{2(1 - \eta)\beta}\|x_A^k - x^*_B\|^2 + 2\gamma_k\mu_C\eta\|x_B^k - x^*\|^2 \\
+ 2\gamma_k\mu_B\|x_B^{k+1} - x^*\|^2.
\]

Thus, combine (C.4) and (C.7) to get

\[
(1 + 2\gamma_k\mu_B)\|x_B^{k+1} - x^*\|^2 + \gamma_k^2\|u_B^{k+1} - u_B^*\|^2 + \left( 1 - \frac{\gamma_k}{2(1 - \eta)\beta} \right)\|x_A^k - x^*_B\|^2 \\
\leq (1 - 2\gamma_k\mu_C\eta)\|x_B^k - x^*\|^2 + \gamma_k^2\|u_B^k - u_B^*\|^2.
\]

Part 2: This follows the exact same reasoning, except we replace Equation (C.6) with the following lower bound:

\[
2\langle x_A^k - x^*, Cx_B^k \rangle = 2\langle x_A^k - x_B^k, Cx_B^k - Cx^* \rangle + 2\langle x_A^k - x_B^k, Cx^* \rangle + 2\langle x_B^k - x^*, Cx_B^k \rangle \\
\geq -\frac{1}{\gamma_k}\|x_A^k - x_B^k\|^2 - \gamma_k\|Cx_B^k - Cx^*\|^2 + 2\langle x_A^k - x^*, Cx^* \rangle \\
\geq -\frac{1}{\gamma_k}\|x_A^k - x_B^k\|^2 - \gamma_k\bar{L}_C^2\|x_B^k - x^*\|^2 + 2\langle x_A^k - x^*, Cx^* \rangle.
\]

We are now ready to prove Theorem 6.1.2.
Theorem 6.1.2. Part 1: The definition of $\gamma_{k+1}$ ensures that

$$\frac{1 + 2 \gamma_k \mu_B}{\gamma_k^2} = \frac{(1 - 2 \gamma_{k+1} \mu \eta)}{\gamma_{k+1}^2}.$$ 

Therefore, by (C.2), the following inequality holds for all $k \geq 0$:

$$\frac{(1 - \gamma_{k+1} \mu \eta)}{\gamma_{k+1}^2} \|x_B^{k+1} - x^*\|^2 + \|u_B^{k+1} - u_B^*\|^2 \leq \frac{(1 - \gamma_k \mu \eta)}{\gamma_k^2} \|x_B^k - x^*\|^2 + \|u_B^k - u_B^*\|^2.$$

Now observe that from Equation (6.1.8), we have $\gamma_k \to 0$ as $k \to \infty$. Therefore,

$$\gamma_k \gamma_{k+1}^2 = \sqrt{1 + 2 \gamma_k \mu_B \gamma_k - \gamma_{k+1} \mu \eta} \to 1$$ 

as $k \to \infty$.

In addition, the sequence $1/\gamma_j$ is increasing:

$$\gamma_k^2 - \gamma_{k+1}^2 = \gamma_k \gamma_{k+1} (2 \gamma_k \mu_B + 2 \gamma_{k+1} \mu \eta) > 0.$$

Thus, we apply the Stolz-Cesàro theorem to compute the following limit:

$$\lim_{k \to \infty} (k + 1) \gamma_k = \lim_{k \to \infty} \frac{k + 1}{\gamma_k} = \lim_{k \to \infty} \frac{(k + 2) - (k + 1)}{\gamma_{k+1} - \gamma_k} = \lim_{k \to \infty} \frac{\gamma_k \gamma_{k+1} + \gamma k}{\gamma_k \gamma_{k+1} + \gamma k}$$

$$= \lim_{k \to \infty} \frac{\gamma_k \gamma_{k+1} (2 \gamma_k \mu_B + 2 \gamma_{k+1} \mu \eta)}{2 \gamma_k \mu_B + 2 \mu \eta} = \frac{1}{\mu \eta + \mu_B}.$$

Thus, we have

$$\|x_B^{k+1} - x^*\|^2 \leq \frac{\gamma_{k+1}^2}{(1 - \gamma_{k+1} \mu \eta)} \left( 1 - \gamma_0 \mu \eta \right) \|x_B^0 - x^*\|^2 + \|u_B^0 - u_B^*\|^2$$

$$= O \left( \frac{1}{(k + 1)^2} \right).$$

Part 2: The proof is nearly identical to the proof of Part 1. The difference is that the definition of $\gamma_{k+1}$ ensures that for all $k \geq 0$, we have

$$\frac{1}{\gamma_{k+1}^2} = \frac{1 + 2 \gamma_k (\mu_B - \gamma_k L^2_C/2)}{\gamma_k^2}.$$
In addition, we have $\gamma_k \to 0$ as $k \to \infty$. The sequence $(1/\gamma_j)_{j \geq 0}$ is also increasing because $\gamma_k < 2L_C^2/\mu_B$ for all $k \geq 0$. Finally we note that $\gamma_k/\gamma_{k+1} \to 1$ as $k \to \infty$. Thus, we apply the Stolz-Cesàro theorem to compute the following limit:

$$
\lim_{k \to \infty} (k + 1)\gamma_k = \lim_{k \to \infty} \frac{k + 1}{\gamma_k} = \lim_{k \to \infty} \frac{(k + 2) - (k + 1)}{\gamma_{k+1} - \gamma_k} = \lim_{k \to \infty} \frac{\gamma_k \gamma_{k+1}}{\gamma_k - \gamma_{k+1}}
$$

$$
= \lim_{k \to \infty} \frac{\gamma_k \gamma_{k+1} (\gamma_{k+1} + \gamma_k)}{\gamma_k^2 - \gamma_{k+1}^2} = \lim_{k \to \infty} \frac{\gamma_k^2}{2} \frac{\gamma_k + \gamma_{k+1}}{\gamma_k \gamma_{k+1}}
$$

$$
= \lim_{k \to \infty} \frac{\gamma_k + \gamma_{k+1}}{\gamma_{k+1}(2\mu_B - \gamma_k L_C^2/2)} = \lim_{k \to \infty} \frac{1 + \gamma_k/\gamma_{k+1}}{2\mu_B - \gamma_k L_C^2} = \frac{1}{\mu_B}.
$$

\[\square\]

### C.2 Derivation of Algorithm 6.2.9

Observe the following identities from Fig. 6.1 and Lemma 6.3.1:

$$
x^k_A - x^{k+1}_B = \gamma (u^{k+1}_B - u^k_B); \quad (C.8a)$$

$$
x^k_B - x^{k+1}_B = \gamma (u^{k+1}_B + Cx^k_B + u^k_A); \quad (C.8b)$$

$$
x^k_B - x^k_A = \gamma (u^k_B + Cx^k_B + u^k_A). \quad (C.8c)
$$

These give us the further subgradient identity:

$$
u^{k+1}_A = u^k_A + (u^{k+1}_A + u^k_B + Cx^k_B) + (C_x^k_B - Cx^{k+1}_B) - (u^k_A + u^{k+1}_B + Cx^k_B)
$$

$$
= u^k_A + \frac{1}{\gamma} (x^{k+1}_B - x^k_A) + (C_x^k_B - Cx^{k+1}_B) + \frac{1}{\gamma} (x^{k+1}_B - x^k_B)
$$

$$
= J_{\frac{1}{\gamma}A^{-1}} \left( u^k_A + \frac{1}{\gamma} (2x^{k+1}_B - x^k_B) + (C_x^k_B - Cx^{k+1}_B) \right),
$$

where the first equality follows from cancellation, the second from (C.8), and the third from the property:

for any $v \in \mathcal{H}$, $u^{k+1}_A = v - \frac{1}{\gamma} x^{k+1}_A$, $(x^{k+1}_A, u^{k+1}_A) \in \text{gra } A \iff u^{k+1}_A = J_{\frac{1}{\gamma}A^{-1}}(v)$,

which follows from the definition of resolvent $J_{\frac{1}{\gamma}A^{-1}}$. In addition,

$$
x^{k+1}_B = x^k_B - \gamma (u^{k+1}_B + Cx^k_B + u^k_A)
$$

$$
= J_{\gamma B} (x^k_B - \gamma Cx^k_B - \gamma u^k_A),
$$

269
where the second equality follows from the property

for any $v \in \mathcal{H}$, $x^{k+1}_B = v - \gamma u^{k+1}_B$, $(x^{k+1}_B, u^{k+1}_B) \in \text{gra } B \iff x^{k+1}_B = J_{\gamma B}(v)$.

Altogether, for all $k \geq 0$, we have

$$x^{k+1}_B = J_{\gamma B}\left(x^k_B - \gamma Cx^k_B - \gamma u^k_A\right),$$

$$u^{k+1}_A = J_{\frac{1}{\gamma} A^{-1}}\left(u^k_A + \frac{1}{\gamma}(2x^{k+1}_B - x^k_B) + (Cx^k_B - Cx^{k+1}_B)\right).$$

Algorithm 6.2.9 is obtained with the change of variable: $x^k \leftarrow x^k_B$ and $y^k \leftarrow u^k_A$.

C.3 Proofs from Section 6.3

of Lemma 6.3.2. Let $x \in \text{zer}(A + B + C)$, that is, $0 \in (A + B + C)x$. Let $u_A \in Ax$ and $u_B \in Bx$ be such that $u_A + u_B + Cx = 0$. In addition, let $z = x + \gamma u_B$. We will show that $z$ is a fixed point of $T$. Then $J_{\gamma B}(z) = x$ and $2J_{\gamma B}(z) - z - \gamma CJ_{\gamma B}(z) = 2x - z - \gamma Cx = x - \gamma Cx - \gamma u_B = x + \gamma u_A$. Thus, $x = J_{\gamma A}(x + \gamma u_A) = J_{\gamma A}(2J_{\gamma B}(z) - z - \gamma CJ_{\gamma B}(z))$. Therefore,

$$Tz = T(x + \gamma u_B)$$

$$= J_{\gamma A}(2J_{\gamma B}(z) - z - \gamma CJ_{\gamma B}(z)) + (I - J_{\gamma B})(z)$$

$$= x + \gamma u_B$$

$$= z.$$

Next, suppose that $z \in \text{Fix } T$. Then there exists $u_B \in B(J_{\gamma B}(z))$ and $u_A \in A(J_{\gamma A}(2J_{\gamma B}(z) - z - \gamma CJ_{\gamma B}(z)))$ such that

$$z = Tz$$

$$= z + J_{\gamma A}(2J_{\gamma B}(z) - z - \gamma CJ_{\gamma B}(z)) - J_{\gamma B}(z)$$

$$= z - \gamma (u_A + u_B + CJ_{\gamma B}(z)).$$

Thus, $x = J_{\gamma A}(2J_{\gamma B}(z) - z - \gamma CJ_{\gamma B}(z)) = J_{\gamma B}(z)$ and $u_A + u_B + Cx = 0$. Therefore, $x = J_{\gamma B}(z) \in \text{zer}(A + B + C)$.
The identity for $\text{Fix}T$ immediately follows from the fixed-point construction process in the first paragraph.

of Lemma 6.3.3. Let $z, w \in \mathcal{H}$. Then

\[
\|S_z - S_w\|^2 = \|U_z - U_w\|^2 + \|T_1 \circ V_z - T_1 \circ V_w\|^2 + 2\langle T_1 \circ V_z - T_1 \circ V_w, U_z - U_w \rangle
\]

\[
\leq \langle U_z - U_w, z - w \rangle + \langle T_1 \circ V_z - T_1 \circ V_w, V_z - V_w \rangle + 2\langle T_1 \circ V_z - T_1 \circ V_w, U_z - U_w \rangle
\]

\[
= \langle U_z - U_w, z - w \rangle + \langle T_1 \circ V_z - T_1 \circ V_w, (2U + V)z - (2U + V)w \rangle
\]

\[
= \langle U_z - U_w, z - w \rangle + \langle T_1 \circ V_z - T_1 \circ V_w, (I - W)z - (I - W)w \rangle
\]

\[
= \langle S_z - S_w, z - w \rangle - \langle T_1 \circ V_z - T_1 \circ V_w, W_z - W_w \rangle
\]

where the inequality follows from the firm nonexpansiveness of $U$ and $T_1$. Then, the result follows from the identity:

\[
\langle S_z - S_w, z - w \rangle = \frac{1}{2}\|z - w\|^2 - \frac{1}{2}\|(I_{\mathcal{H}} - S)z - (I_{\mathcal{H}} - S)w\|^2 + \frac{1}{2}\|S_z - S_w\|^2.
\]

C.4 Proofs for convergence rate analysis

We now recall a lower bound property for convex functions that are strongly convex and Lipschitz differentiable. The first bound is a consequence of [11, Theorem 18.15] and the second bound is a combination of [11, Theorem 18.15] and [94, Theorem 2.1.12].

Proposition C.2. Suppose that $f : \mathcal{H} \to (0, \infty]$ is $\mu$-strongly convex and $(1/\beta)$-Lipschitz differentiable. For all $x, y \in \text{dom}(f)$, let

\[
S_f(x, y) := \max\left\{\frac{\beta}{2}\|\nabla f(x) - \nabla f(y)\|^2, \frac{\mu}{2}\|x - y\|^2\right\},
\]

\[
Q_f(x, y) := \max\left\{2S_f(x, y), \frac{\mu}{(\mu\beta + 1)}\|x - y\|^2 + \frac{\beta}{(\mu\beta + 1)}\|\nabla f(x) - \nabla f(y)\|^2\right\}.
\]
Then for all \(x, y \in \text{dom}(f)\), we have
\[
f(x) - f(y) - (x - y, \nabla f(y)) \geq S_f(x, y); \quad (C.11)
\]
\[
(\nabla f(x) - \nabla f(y), x - y) \geq Q_f(x, y). \quad (C.12)
\]

Similarly, if \(A : \mathcal{H} \to \mathcal{H}\) is \(\mu\)-strongly monotone and \(\beta\)-cocoercive, we let
\[
Q_A(x, y) = \max \{\mu\|x - y\|^2, \beta\|Ax - Ay\|^2\}
\]
for all \(x, y \in \text{dom}(A)\). Then for all \(x, y \in \text{dom}(A)\), we have
\[
(Ax - Ay, x - y) \geq Q_A(x, y).
\]

We follow the convention that every function \(f\) is \(\mu_f \geq 0\) strongly convex and \((1/\beta_f) \geq 0\) Lipschitz where we allow the possibility that \(\beta_f = \mu_f = 0\). With this notation, the results of Proposition C.2 continue hold for all \(f\). We follow the same convention for monotone operators. In particular, every monotone operator \(A : \mathcal{H} \to 2^{\mathcal{H}}\) is \(\mu_A\)-strongly monotone and \(\beta_A\)-cocoercive where \(\mu_A \geq 0\) and \(\beta_A \geq 0\). Finally, we follow convention that \(Q_{\partial f} := Q_f\).

Note that we could extend our definition of \(Q_A(\cdot, \cdot)\) (or \(Q_f(\cdot, \cdot)\)) to the case where \(A\) is merely strongly monotone in a subset of the coordinates of \(\mathcal{H}\) (which is then assumed to be a product space). This extension is straightforward, though slightly messy. Thus, we omit this extension.

The following identity will be applied repeatedly:

**Proposition C.3.** Let \(z \in \mathcal{H}\), let \(z^*\) be a fixed point of \(T\), let \(\gamma > 0\), let \(\lambda > 0\), and let \(z^+ = (1 - \lambda)z + \lambda Tz\). Then
\[
2\gamma \lambda (x_B - x_A, u_B^* + Cx^*) + 2\gamma \lambda Q_A(x_A, x^*) + 2\gamma \lambda Q_B(x_B, x^*) + 2\gamma \lambda Q_C(x_B, x^*)
\]
\[
\leq 2\gamma \lambda (x_A - x^*, u_A) + 2\gamma \lambda (x_B - x^*, u_B + Cx_B) \quad (C.13)
\]
\[
= \|z - x^*\|^2 - \|z^+ - x^*\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z - z^+\|^2 + 2\gamma \langle z - z^+, Cx_B\rangle \quad (C.14)
\]
\[
= \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z - z^+\|^2 + 2\gamma \langle z - z^+, Cx_B + u_B^*\rangle \quad (C.15)
\]
where \( x_A \in \text{dom}(A), x_B \in \text{dom}(B), u_B \in Bx_B \) and \( u_A \in Au_A \) are defined in Lemma 6.3.1 and Equation (C.14) holds for all \( x^* \in \mathcal{H} \), while Equations (C.13) and (C.15) hold when \( x^* = J_{\gamma B}(z^*) \) and \( u_B^* = \frac{1}{\gamma}(z^* - x^*) \). In particular, when \( x^* = J_{\gamma B}(z^*) \), we have

\[
\|z^* - z^*\|^2 + \frac{2}{\lambda} \|z - z^+\|^2 + 2\gamma\lambda Q_A(x_A, x^*) + 2\gamma\lambda Q_B(x_B, x^*) + 2\gamma\lambda Q_C(x_B, x^*)
\leq \|z - z^*\|^2 + 2\gamma\langle z - z^+, C(x_B) - C(x^*) \rangle.
\]

**(Proof.** First we show inequality (C.13): Let \( u_A^* \in Ax^* \) and \( u_B^* \in Bx^* \) be such that \( u_A^* + u_B^* + Cx^* = 0 \). Then

\[
2\gamma\lambda\langle x_A - x^*, u_A \rangle + 2\gamma\lambda\langle x_B - x^*, u_B + Cx_B \rangle \\
\geq 2\gamma\lambda\langle x_A - x^*, u_A^* \rangle + 2\gamma\lambda\langle x_B - x^*, u_B^* + Cx^* \rangle \\
+ 2\gamma\lambda Q_A(x_A, x^*) + 2\gamma\lambda Q_B(x_B, x^*) + 2\gamma\lambda Q_C(x_B, x^*) \\
= \gamma\lambda\langle x_A - x^*, u_A^* + u_B^* + Cx^* \rangle + 2\gamma\lambda\langle x_B - x_A, u_B^* + Cx^* \rangle \\
+ 2\gamma\lambda Q_A(x_A, x^*) + 2\gamma\lambda Q_B(x_B, x^*) + 2\gamma\lambda Q_C(x_B, x^*) \\
= 2\gamma\lambda\langle x_B - x_A, u_B^* + Cx^* \rangle + 2\gamma\lambda Q_A(x_A, x^*) + 2\gamma\lambda Q_B(x_B, x^*) + 2\gamma\lambda Q_C(x_B, x^*).
\]

Now we show Equation (C.14):

\[
2\lambda\gamma\langle x_A - x^*, u_A \rangle + 2\gamma\lambda\langle x_B - x^*, u_B + Cx_B \rangle \\
= 2\gamma\lambda\langle x_A - x_B, u_A \rangle + 2\gamma\lambda\langle x_B - x^*, u_A + u_B + Cx_B \rangle \\
= 2\lambda\langle x_A - x_B, \gamma u_A \rangle + 2\lambda\langle x_B - x^*, x_B - x_A \rangle \\
= 2\lambda\langle x_B - \gamma u_A - x^*, x_B - x_A \rangle \\
= 2\langle z + (x_B - z - \gamma u_A) - x^*, z - z^+ \rangle \\
= 2\langle z - \gamma(u_B + u_A + Cx_B) - x^*, z - z^+ \rangle + 2\gamma\langle z - z^+, Cx_B \rangle \\
= 2\langle z - \frac{1}{\lambda}(z - z^+) - x^*, z - z^+ \rangle + 2\gamma\langle z - z^+, Cx_B \rangle \\
= 2\langle z - x^*, z - z^+ \rangle - \frac{2}{\lambda}\|z - z^+\|^2 + 2\gamma\langle z - z^+, Cx_B \rangle \\
\overset{(6.1.12)}{=} \|z - x^*\|^2 - \|z^+ - z^*\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z - z^+\|^2 + 2\gamma\langle z - z^+, Cx_B \rangle.
\]
Now assume that \( x^* = J_{zB}(z^*) \) and show Equation (C.15):

\[
2\lambda\gamma\langle x_A - x^*, u_A \rangle + 2\gamma\lambda\langle x_B - x^*, u_B + Cx_B \rangle \\
= 2\langle z - x^*, z - z^+ \rangle - \frac{2}{\lambda}\|z - z^+\|^2 + 2\gamma\langle z - z^+, Cx_B \rangle \\
= 2\langle z - z^*, z - z^+ \rangle - \frac{2}{\lambda}\|z - z^+\|^2 + 2\gamma\langle z - z^+, Cx_B + u_B^* \rangle \\
\overset{(6.1.12)}{=} \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z - z^+\|^2 + 2\gamma\langle z - z^+, Cx_B + u_B^* \rangle.
\]

Equation (C.16) follows from rearranging the above inequalities. \( \square \)

**Corollary C.1 (Function value bounds).** Assume the notation of Proposition C.3. Let \( f, g, \) and \( h \) be closed, proper and convex functions from \( \mathcal{H} \) to \( (-\infty, \infty] \). Suppose that \( h \) is \((1/\beta)\)-Lipschitz differentiable. Suppose that \( A = \partial f, \) \( B = \partial g, \) and \( C = \nabla h. \) Then if \( x^* = \text{prox}_{\gamma g}(z^*), \) \( \nabla g(x^*) = (1/\gamma)(z^* - x^*), \) and \( \nabla f(x^*) \in \partial f(x^*) \) and \( \nabla g(x^*) \in \partial g(x^*) \) are such that \( \nabla h(x^*) + \nabla g(x^*) + \nabla f(x^*) = 0, \) we have

\[
2\gamma\langle z - z^+, \nabla g(x^*) + \nabla h(x^*) \rangle + 4\gamma\lambda S_f(x_f, x^*) + 4\gamma\lambda S_g(x_g, x^*) + 4\gamma\lambda S_h(x_g, x^*) \\
\leq 2\gamma\lambda(f(x_f) + g(x_g) + h(x_g) - (f + g + h)(x^*) + S_f(x_f, x^*) + S_g(x_g, x^*) + S_h(x_g, x^*)) \\
\leq \|z - x^*\|^2 - \|z^+ - x^*\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z - z^+\|^2 + 2\gamma\langle z - z^+, \nabla h(x_B) \rangle \\
= \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z - z^+\|^2 + 2\gamma\langle z - z^+, \nabla h(x_B) + \nabla g(x^*) \rangle.
\]

where \( x_f \in \text{dom}(f), x_g \in \text{dom}(g) \) are defined in Lemma 6.3.1 and Equation (C.18) holds for all \( x^* \in \mathcal{H}, \) while Equations (C.17) and (C.19) hold when \( x^* = \text{prox}_{\gamma g}(z^*) \) and \( \nabla g(x^*) = \frac{1}{\gamma}(z^* - x^*). \) In particular, when \( x^* = \text{prox}_{\gamma g}(z^*), \) we have

\[
\|z^+ - z^*\|^2 + \left(\frac{2}{\lambda} - 1\right)\|z - z^+\|^2 + 4\gamma\lambda S_f(x_f, x^*) + 4\gamma\lambda S_g(x_g, x^*) + 4\gamma\lambda S_h(x_g, x^*) \\
\leq \|z - z^*\|^2 + 2\gamma\langle z - z^+, \nabla h(x_B) - \nabla g(x^*) \rangle.
\]
Proof. Equation (C.19) is a direct consequence of Proposition C.3 together with the inequalities:

\[ f(x_f) + g(x_g) + h(x_g) - (f + g + h)(x^*) \leq \langle x_f - x^*, \nabla f(x_f) \rangle + \langle x_g - x^*, \nabla g(x_g) + \nabla h(x_g) \rangle - S_f(x_f, x^*) - S_g(x_g, x^*) - S_h(x_g, x^*); \]

\[ f(x_f) + g(x_g) + h(x_g) - (f + g + h)(x^*) \geq \langle x_f - x^*, \tilde{\nabla} f(x^*) \rangle + \langle x_g - x^*, \tilde{\nabla} g(x^*) + \nabla h(x^*) \rangle + S_f(x_f, x^*) + S_g(x_g, x^*) + S_h(x_g, x^*) \]

where we use that \( x_g - x_f = (1/\lambda)(z - z^+) \) (see Lemma 6.3.1.)

Equation (C.20) is a consequence of the Equation (C.16). \( \square \)

Corollary C.2 (Subdifferentiable + monotone model variational inequality bounds).

Assume the notation of Proposition C.3. Let \( f, g, \) and \( h \) be closed, proper and convex functions from \( \mathcal{H} \) to \( (-\infty, \infty] \), and let \( \nabla h \) be \((1/\beta_h)\)-Lipschitz. Let \( \overline{A}, \overline{B} \) and \( C \) be monotone operators on \( \mathcal{H} \), and let \( \overline{C} \) be \( \beta_C \)-cocoercive. Suppose that \( A = \partial f + \overline{A}, B = \partial g + \overline{B} \), and \( C = \nabla h + \overline{C} \). Let \( \tilde{\nabla} f(x_A) + u_\overline{A} = u_A \) where \( \tilde{\nabla} f(x_A) \in \partial f(x_A) \) and \( u_\overline{A} \in \overline{A} x_A \). Likewise let \( \tilde{\nabla} g(x_B) + u_\overline{B} = u_B \) where \( \tilde{\nabla} g(x_B) \in \partial g(x_B) \) and \( u_\overline{B} \in \overline{A} x_A \). Then for all \( x \in \text{dom}(f) \cap \text{dom}(g), \) we have

\[ 2\gamma \lambda \left( f(x_A) + g(x_B) + h(x_B) - (f + g + h)(x) + S_f(x_f, x) + S_g(x_g, x) + S_h(x_g, x) \right. \]

\[ \left. + \langle x_A - x, u_\overline{A} \rangle + \langle x_B - x, u_\overline{B} + \overline{C} x_B \rangle \right) \]

\[ \leq \| z - x \|^2 - \| z^+ - x \|^2 + \left( 1 - \frac{2}{\lambda} \right) \| z - z^+ \|^2 + 2\gamma \langle z - z^+, \tilde{\nabla} h(x_B) + \overline{C} x_B \rangle \quad \text{(C.21)} \]

Proof. Equation (C.19) is a direct consequence of Proposition C.3 together with the fol-
f(x_f) + g(x_g) + h(x_g) - (f + g + h)(x)
\leq \langle x_f - x, \nabla f(x_f) \rangle + \langle x_g - x, \nabla g(x_g) + \nabla h(x_g) \rangle
- S_f(x_f, x^*) - S_g(x_g, x^*) - S_h(x_g, x^*).

C.4.1 General case: convergence rates of upper and lower bounds

We will prove the most general rates by showing how fast the upper and lower bounds in Proposition C.3 converge. Then we will deduce convergence rates. Thus, in this section we set

$$\kappa_1^k(\lambda, x) = \|z^k - x\|^2 - \|z^{k+1} - x^*\|^2$$
$$+ \left(1 - \frac{2}{\lambda}\right)\|z^k - z^{k+1}\|^2 + 2\gamma\langle z^k - z^{k+1}, Cx_B^k \rangle$$
$$\kappa_2^k(\lambda, x^*) = \|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2$$
$$+ \left(1 - \frac{2}{\lambda}\right)\|z^k - z^{k+1}\|^2 + 2\gamma\langle z^k - z^{k+1}, Cx_B^k - Cx^* \rangle$$

where $\lambda > 0$, $z^*$ is a fixed point of $T$, $x^* = J_{\gamma B}(z^*)$, and $x \in H$.

**Theorem C.1** (Nonergodic convergence rates of bounds). Let $(z^j)_{j \geq 0}$ be generated by Equation (6.1.4) with $\varepsilon \in (0, 1), \gamma \in (0, 2\beta \varepsilon), \alpha = 1/(2 - \varepsilon) < 2\beta/(4\beta - \gamma)$, and $(\lambda_j)_{j \geq 0} \subseteq (0, 1/\alpha)$. Let $z^*$ be a fixed point of $T$, let $x^* = J_{\gamma B}(z^*)$, and let $x \in H$. Assume that $\overline{\tau} := \inf_{j \geq 0} \lambda_j (1 - \alpha \lambda_j)/\alpha$. Then for all $k \geq 0$,

$$\kappa_1^k(1, x) \leq \frac{2(\|z^* - x\| + (1 + \gamma/\beta)\|z^0 - z^*\| + \|C x^*\|)\|z^0 - z^*\|}{\sqrt{\overline{\tau}(k + 1)}}$$

$$|\kappa_1^k(1, x)| = o\left(\frac{1 + \|x\|}{\sqrt{k + 1}}\right);$$

(C.22)
and

\[ \kappa_2^k(1, x^*) \leq \frac{2(1 + \gamma/\beta)\|z^0 - z^*\|^2}{\sqrt{T(k + 1)}}, \]

\[ |\kappa_2^k(1, x^*)| = o \left( \frac{1}{\sqrt{k + 1}} \right). \quad (C.23) \]

We also have the following lower bound:

\[ \langle x_B^k - x_A^k, u_B^* + Cx^* \rangle \geq -\|z^0 - z^*\|\|u_B^* + Cx^*\| \frac{1}{\sqrt{T(k + 1)}}, \]

\[ |\langle x_B^k - x_A^k, u_B^* + Cx^* \rangle| = o \left( \frac{1}{\sqrt{k + 1}} \right). \quad (C.24) \]

**Proof.** Fix \( k \geq 0 \). Observe that

\[ \|Cx_B^k\| \leq \|Cx_B^k - Cx^*\| + \|Cx^*\| \leq \frac{1}{\beta}\|x_B^k - x^*\| + \|Cx^*\| \leq \frac{1}{\beta}\|z^k - z^0\| + \|Cx^*\| \]

by the \((1/\beta)\)-Lipschitz continuity of \( C \), the nonexpansiveness of \( J_{\gamma B} \), and the monotonicity of the sequence \((\|z^j - z^*\|)_{j \geq 0}\) (see Part 1 of Theorem 6.3.1). Thus,

\[ |\kappa_1^k(1, x)| \overset{(6.1.12)}{=} \left| 2(z^{k+1} - x, z^k - z^{k+1}) + 2\gamma(z^k - z^{k+1}, Cx_B^k) \right| \]

\[ \leq 2\|z^{k+1} - x\|\|z^0 - z^*\| + (2\gamma/\beta)\|z^0 - z^*\|^2 + 2\gamma\|Cx^*\|\|z^0 - z^*\| \sqrt{T(k + 1)} \]

\[ \leq \frac{(2\|z^* - x\| + (2 + 2\gamma/\beta)\|z^0 - z^*\|^2 + 2\gamma\|Cx^*\|)|z^0 - z^*|}{\sqrt{T(k + 1)}} \]

where the bound in the second inequality follows from Cauchy-Schwarz and the upper bound in Part 7 of Theorem 6.3.1, and the last inequality follows because \( \|z^{k+1} - x\| \leq \|z^k - z^*\| + \|z^* - x\| \) (see Part 1 of Theorem 6.3.1). The little-o rate follows because \( \|z^k - z^{k+1}\| = o \left( \frac{1}{\sqrt{k + 1}} \right) \) by Part 7 of Theorem 6.3.1.

The proof of Equation (C.23) follows nearly the same reasoning as the proof of Equation (C.22). Thus, we omit the proof.

Next, because \( x_B^k - x_A^k = z^k - Tz^k \) (see Lemma 6.3.1), we have

\[ |\langle z^k - Tz^k, u_B^* + Cx_B^* \rangle| \leq \frac{\|z^0 - z^*\|\|u_B^* + Cx^*\|}{\sqrt{T(k + 1)}}. \]
by Part 7 of Theorem 6.3.1. Similarly The little-o rate follows because \( \|z^k - z^{k+1}\| = o\left(1/\sqrt{k + 1}\right) \) by Part 7 of Theorem 6.3.1.

We now prove two ergodic results.

**Theorem C.2** (Ergodic convergence rates of bounds for Equation (6.1.6)). Let \((z^i)_{j \geq 0}\) be generated by Equation (6.1.4) with \(\varepsilon \in (0, 1), \gamma \in (0, 2\beta\varepsilon), \alpha = 1/(2 - \varepsilon) < 2\beta/(4\beta - \gamma)\), and \((\lambda_j)_{j \geq 0} \subseteq (0, 1/\alpha]\). Let \(z^*\) be a fixed point of \(T\), let \(x^* = J_\gamma B(z^*)\), and let \(x \in \mathcal{H}\). Then for all \(k \geq 0\),

\[
\frac{1}{\sum_{i=0}^{k} \lambda_i} \sum_{i=0}^{k} \kappa^i_1(\lambda_i, x) \leq \frac{\|z^0 - x\|^2 + \frac{\gamma}{(2\beta\varepsilon - \gamma)}\|z^0 - z^*\|^2 + 4\gamma\|z^0 - z^*\||Cx^*|}{\sum_{i=0}^{k} \lambda_i} \quad \text{(C.25)}
\]

\[
\frac{1}{\sum_{i=0}^{k} \lambda_i} \sum_{i=0}^{k} \kappa^i_2(\lambda_i, x^*) \leq \frac{\left(1 + \frac{\gamma}{(2\beta\varepsilon - \gamma)}\right)\|z^0 - z^*\|^2}{\sum_{i=0}^{k} \lambda_i}. \quad \text{(C.26)}
\]

We also have the following lower bound:

\[
\frac{1}{\sum_{i=0}^{k} \lambda_i} \sum_{i=0}^{k} \lambda_i \langle x^i_B - x^i_A, u^*_B + Cx^* \rangle \geq \frac{-2\|z^0 - z^*\||u^*_B + Cx^*|}{\sum_{i=0}^{k} \lambda_i}. \quad \text{(C.27)}
\]

In addition, the following feasibility bound holds:

\[
\left\| \frac{1}{\sum_{i=0}^{k} \lambda_i} \sum_{i=0}^{k} \lambda_i \langle x^i_B - x^i_A \rangle \right\| \leq \frac{2\|z^0 - z^*\|}{\sum_{i=0}^{k} \lambda_i}. \quad \text{(C.28)}
\]

**Proof.** Fix \(k \geq 0\). We first prove the feasibility bound:

\[
\left\| \frac{1}{\sum_{i=0}^{k} \lambda_i} \sum_{i=0}^{k} \lambda_i \langle x^i_B - x^i_A \rangle \right\| \leq \frac{1}{\sum_{i=0}^{k} \lambda_i} \left\| \sum_{i=0}^{k} \langle z^i - z^{i+1} \rangle \right\| = \frac{\|z^0 - z^{k+1}\|}{\sum_{i=0}^{k} \lambda_i} \leq \frac{2\|z^0 - z^*\|}{\sum_{i=0}^{k} \lambda_i},
\]

where the last inequality follows from \(\|z^0 - z^{k+1}\| \leq \|z^0 - z^*\| + \|z^{k+1} - z^*\| \leq 2\|z^0 - z^*\|\).

Let \(\eta_k = 2/\lambda_k - 1\). Note that \(\eta_k > 0\), by assumption. In addition, \(1/\eta_k = \lambda_k/(2 - \lambda_k) \leq \lambda_k/\varepsilon\). Thus, we have

\[
2\gamma\langle z^k - z^{k+1}, Cx^k_B \rangle = 2\gamma\langle z^k - z^{k+1}, Cx^k_B - Cx^* \rangle + 2\gamma\langle z^k - z^{k+1}, Cx^* \rangle 
\]

\[
\leq \eta_k\|z^k - z^{k+1}\|^2 + \frac{\gamma^2}{\eta_k}\|Cx^k_B - Cx^*\|^2 + 2\gamma\langle z^k - z^{k+1}, Cx^* \rangle \quad \text{(C.29)}
\]

278
Thus, for all \( k \geq 0 \), we have

\[
\sum_{i=0}^{k} \kappa_i^1(\lambda_i, x) \leq \|z^0 - x\|^2 + \sum_{i=0}^{k} \left( \eta_i \|z^{i+1} - z^i\|^2 \right) + \sum_{i=0}^{k} \left( \eta_i \|z^{i+1} - z^i\|^2 \right) + \sum_{i=0}^{k} \left( 2\gamma \langle z^i - z^{i+1}, Cx_B^i \rangle \right) \]

\[(\text{C.29}) \leq \|z^0 - x\|^2 + \sum_{i=0}^{k} \left( \frac{\gamma^2 \lambda_i}{\varepsilon} \|Cx^i_B - Cx^*\|^2 \right) + \sum_{i=0}^{k} \left( 2\gamma \langle z^i - z^{i+1}, Cx^* \rangle \right) \]

\[
\leq \|z^0 - x\|^2 + \frac{\gamma^2}{\varepsilon(2\beta - \gamma/\varepsilon)} \|z^0 - z^*\|^2 + 2\gamma \langle z^0 - z^{k+1}, Cx^* \rangle \]

\[
\leq \|z^0 - x\|^2 + \frac{\gamma}{(2\beta - \gamma)} \|z^0 - z^*\|^2 + 4\gamma \|z^0 - z^*\|\|Cx^*\|.
\]

where the third inequality follows from Part 4 of Theorem 6.3.1 and the fourth inequality follows because \( \|z^0 - z^{k+1}\| \leq \|z^0 - z^*\| + \|z^{k+1} - z^*\| \leq 2\|z^0 - z^*\| \).

The proof of Equation (C.26) follows nearly the same reasoning as the proof of Equation (C.25). Thus, we omit the proof.

Finally, Equation (C.27) follows directly from Cauchy Schwarz and Equation (C.28).

\[\square\]

**Theorem C.3** (Ergodic convergence rates of bounds for Equation (6.1.7)). Let \((z^j)_{j \geq 0}\) be generated by Equation (6.1.4) with \(\varepsilon \in (0, 1), \gamma \in (0, 2\beta \varepsilon), \alpha = 1/(2 - \varepsilon) < 2\beta/(4\beta - \gamma),\) and \(\lambda_j \equiv \lambda \subseteq (0, 1/\alpha] \). Let \(z^*\) be a fixed point of \(T\), let \(x^* = J_{\varepsilon B}(z^*)\), and let \(x \in \mathcal{H}\). Then for all \( k \geq 0 \),

\[
\frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} (i+1) \kappa_1^i(\lambda, x) \leq 2 \left( \frac{2\|z^* - x\|^2 + 2 + \frac{\gamma}{(2\beta - \gamma/\varepsilon)} \|z^0 - z^*\|^2 + 10\|z^0 - z^*\|\|Cx^*\|}{k+1} \right); \tag{C.30}
\]

and

\[
\frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} (i+1) \kappa_2^i(\lambda, x^*) \leq \frac{2 \left( 1 + \frac{\gamma}{(2\beta - \gamma/\varepsilon)} \right) \|z^0 - z^*\|^2}{k+1}. \tag{C.31}
\]

We also have the following lower bound:

\[
\frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} (i+1) \langle x_B^i - x_A^i, u_B^i + Cx^* \rangle \geq \frac{-5\|z^0 - z^*\|}{\lambda(k+1)}. \tag{C.32}
\]
In addition, the following feasibility bound holds:

\[
\left\| \frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} (i+1)(x_B^i - x_A^i) \right\| \leq \frac{5\|z^0 - z^*\|}{\lambda(k+1)}.
\] (C.33)

**Proof.** Fix \( k \geq 0 \). We first prove the feasibility bound:

\[
\left\| \frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} (i+1)(x_B^i - x_A^i) \right\| = \frac{1}{\lambda} \left\| \frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} (i+1)(z^i - z^{i+1}) \right\|
\]

\[
\leq \left\| \frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} ((z^{i+1} - z^*) + (i+1)(z^i - z^*) - (i+2)(z^{i+1} - z^*)) \right\|
\]

\[
\leq \left\| \frac{2}{(k+1)(k+2)} \left( \sum_{i=0}^{k} (z^{i+1} - z^*) + (z^0 - z^*) - (k+2)(z^{k+1} - z^*) \right) \right\|
\]

\[
\leq \frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} \|z^{i+1} - z^*\| + \frac{2}{(k+1)(k+2)} \|z^0 - z^*\| + \frac{2}{(k+1)} \|z^{k+1} - z^*\|
\]

\[
\leq \frac{2\|z^0 - z^*\|}{(k+2)} + \frac{2\|z^0 - z^*\|}{(k+1)(k+2)} + \frac{2\|z^0 - z^*\|}{(k+1)}
\]

\[
\leq \frac{2\|z^0 - z^*\| (2 + \frac{1}{k+2})}{(k+1)} \leq \frac{5\|z^0 - z^*\|}{k+1}
\] (C.34)

where we use the bound \( \|z^k - z^*\| \leq \|z^0 - z^*\| \) for all \( k \geq 0 \) (see Part 1 of theorem 6.3.1).

The bound then follows because \( \lambda(x_B^k - x_A^k) = z^k - z^{k+1} \) for all \( k \geq 0 \) (Lemma 6.3.1).
We proceed as in the proof of Theorem C.2 (which is where \( \eta_i := 2/\lambda_i - 1 \) is defined):

\[
\frac{2}{(k + 2)(k + 1)} \sum_{i=0}^{k} (i + 1) \kappa_i^i(\lambda, x)
\]

\[
= \frac{2}{(k + 2)(k + 1)} \sum_{i=0}^{k} \left( (i + 1)\|z^i - x\|^2 - (i + 1)\|z^{i+1} - x\|^2 \right)
\]

\[
+ (i + 1) \left( -\eta_i\|z^{i+1} - z^i\|^2 + 2\gamma\langle z^i - z^{i+1}, Cx^i_B \rangle \right)
\]

\[
\leq \frac{2}{(k + 2)(k + 1)} \sum_{i=0}^{k} \left( \|z^{i+1} - x\|^2 + (i + 1)\|z^i - x\|^2 - (i + 2)\|z^{i+1} - x\|^2 \right)
\]

\[
+ (i + 1) \left( \frac{\gamma^2 \lambda}{\varepsilon}\|Cx^i_B - Cx^*\|^2 + 2\gamma\langle z^i - z^{i+1}, Cx^* \rangle \right)
\]

\[
\leq \frac{2}{(k + 2)(k + 1)} \sum_{i=0}^{k} \|z^i - x\|^2 + \frac{2}{k + 2} \sum_{i=0}^{k} \frac{\gamma^2 \lambda}{\varepsilon}\|Cx^i_B - Cx^*\|^2
\]

\[
+ \frac{2}{(k + 2)(k + 1)} \sum_{i=0}^{k} 2\gamma(i + 1)\langle z^i - z^{i+1}, Cx^* \rangle
\]

\[
\leq \frac{2}{(k + 2)(k + 1)} \sum_{i=0}^{k} \left( 2\|z^i - z^*\|^2 + 2\|z^* - x\|^2 \right) + \frac{2\gamma}{(2\beta\varepsilon - \gamma)} \|z^0 - z^*\|^2
\]

\[
+ \frac{20\gamma\|z^0 - z^*\||Cx^*|}{k + 1}
\]

\[
\leq \frac{2}{k + 1} \left( 2\|z^* - x\|^2 + \left( 2 + \frac{\gamma}{(2\beta\varepsilon - \gamma)} \right) \|z^0 - z^*\|^2 + 10\|z^0 - z^*\||Cx^*| \right)
\]

The proof of Equation (C.31) follows nearly the same reasoning as the proof of Equation (C.30). Thus, we omit the proof.

Finally, Equation (C.32) follows directly from Cauchy Schwarz and Equation (C.33).

\[\square\]

### C.4.2 General case: Rates of function values and variational inequalities

In this section, we use the convergence rates of the upper and lower bounds derived in Theorems C.1, C.2, and C.3 to deduce convergence rates function values and variational
inequalities. All of the convergence rates have the following orders:

Nonergodic: \( o \left( \frac{1}{\sqrt{k+1}} \right) \) and Ergodic: \( O \left( \frac{1}{k+1} \right) \).

We work with three model problems.

- **Most general:** \( A = \partial f + \overline{A} \), \( B = \partial g + \overline{B} \) and \( C = \nabla h + \overline{C} \) where \( f, g \) and \( h \) are functions and \( \overline{A}, \overline{B} \) and \( \overline{C} \) are monotone operators. See Corollary C.2 for our assumptions about this case, and see Corollary C.4 for the nonergodic convergence rate of the variational inequality associated to this problem. Note that for variational inequalities, only upper bounds are important, because we only wish to make certain quantities negative.

- **Subdifferential + Skew:** We use the same set up as above, except we assume that \( \overline{A} \) and \( \overline{B} \) are skew linear mappings (i.e., \( A^* = -A \) and \( B^* = -B \)) and \( \overline{C} = 0 \). See Corollaries C.6 and C.8 for the ergodic convergence rate of the variational inequality associated to this problem. This inclusion problem arises in primal-dual operator-splitting algorithms.

- **Functions:** We assume that \( \overline{A} = \overline{B} = \overline{C} \equiv 0 \). See Corollary C.3 for the nonergodic convergence rate and see Corollaries C.5 and C.7 for the ergodic convergence rates of the function values associated to our method.

Note that by Theorem 2.7.2 of Chapter 2, all of the convergence rates below are sharp (in terms of order, but not necessarily in terms of constants). In addition, they generalize some of the known convergence rates provided in Chapters 2, 4, and 5 for Douglas-Rachford splitting, forward-Douglas-Rachford splitting, and the primal-dual forward-backward splitting, Douglas-Rachford splitting, and the proximal-point algorithms.

The following fact will be used several times:

**Lemma C.1.** Suppose that \( (z^j)_{j \geq 0} \) is generated by Equation (6.1.4) and \( \gamma \in (0, 2\beta) \). Let \( z^* \) be a fixed point of \( T \) and let \( x^* = J_{\gamma B}(z^*) \). Then \( (x^j_A)_{j \geq 0} \) and \( (x^j_B)_{j \geq 0} \) are contained within the closed ball \( B(x^*, (1 + \gamma/\beta)\|z^0 - z^*\|) \).
Proof. Fix $k \geq 0$. Observe that

$$\|x^k_B - x^*\| = \|J_{\gamma B}(z^k) - J_{\gamma B}(z^*)\| \leq \|z^k - z^*\| \leq \|z^0 - z^*\|$$

by Part 1 of Theorem 6.3.1. Similarly,

$$\|x^k_A - x^*\| \leq \|\text{refl}_{\gamma B}(z^k) - \text{refl}_{\gamma B}(z^*) + \gamma Cx^* - \gamma CX_B\|$$

$$\leq \|z^k - z^*\| + \gamma \|z^k - z^*\| \leq \left(1 + \frac{\gamma}{\beta}\right) \|z^0 - z^*\|.$$

\[\square\]

**Corollary C.3** (Nonergodic convergence of function values). Suppose that $(z^j)_{j \geq 0}$ is generated by Equation (6.1.4), with $A = \partial f, B = \partial g$ and $C = \nabla h$. Let the assumptions be as in Theorem C.1. Then the following convergence rates hold:

1. For all $k \geq 0$, we have

$$-\|z^0 - z^*\|\|u^*_B + CX^*\| \leq f(x^k_B) + g(x^k_g) + h(x^k_g) - (f + g + h)(x^*)$$

$$\leq \frac{(\|z^* - x^*\| + (1 + \gamma/\beta)\|z^0 - z^*\| + \gamma\|\nabla h(x^*)\|)\|z^0 - z^*\|}{\gamma \sqrt{\tau(k + 1)}}$$

and

$$|f(x^k_f) + g(x^k_g) - (f + g + h)(x^*)| = o\left(\frac{1}{\sqrt{k + 1}}\right).$$

2. Suppose that $f$ is $L$-Lipschitz continuous on $B(x^*, (1 + \gamma/\beta)\|z^0 - z^*\|)$. Then the following convergence rate holds:

$$0 \leq f(x^k_g) + g(x^k_g) + h(x^k_g) - (f + g + h)(x^*)$$

$$\leq \frac{(\|z^* - x^*\| + (1 + \gamma/\beta)\|z^0 - z^*\| + \gamma\|\nabla h(x^*)\|)\|z^0 - z^*\|}{\gamma \sqrt{\tau(k + 1)}}$$

and

$$0 \leq f(x^k_f) + g(x^k_g) + h(x^k_g) - (f + g + h)(x^*) = o\left(\frac{1}{\sqrt{k + 1}}\right).$$
Proof. Fix $k \geq 0$.

Part 1: By Corollary C.1, we have
\[
\langle x^k_B - x^k_A, u^*_B + Cx^* \rangle \leq f(x^k_f) + g(x^k_g) + h(x^k_g) - (f + g + h)(x^*) \leq \frac{1}{2\gamma} \kappa^k(1,x^*)
\]
Thus, the convergence rates follow directly from Theorem C.1.

Part 2: Note that $f(x^k_f) - f(x^k_g) \leq L\|x^k_f - x^k_g\|$ by Lemma C.1. Because $x_f - x_g = z^k - Tz^k$, we have
\[
f(x^k_f) + g(x^k_g) + h(x^k_g) - (f + g + h)(x^*)
\]
\[
\leq f(x^k_f) + g(x^k_g) + h(x^k_g) - (f + g + h)(x^*) + L\|x^k_f - x^k_g\|
\]
\[
\leq f(x^k_f) + g(x^k_g) + h(x^k_g) - (f + g + h)(x^*) + \frac{\|z^0 - z^*\|}{\sqrt{\tau(k+1)}}.
\]
Thus, the rate follows by Part 1. \hfill \Box

Corollary C.4 (Nonergodic convergence of variational inequalities). Suppose that $(z^j)_{j \geq 0}$ is generated by Equation (6.1.4), with $A = \partial f + \overline{A}$, $B = \partial g + \overline{B}$ and $C = \nabla h + \overline{C}$ as in Corollary C.2. Let the assumptions be as in Theorem C.1. Then the following convergence rates hold:

1. For all $k \geq 0$ and $x \in \text{dom}(f) \cap \text{dom}(g)$, we have
\[
f(x^k_A) + g(x^k_B) + h(x^k_B) - (f + g + h)(x) + \langle x^k_A - x, u^k_A \rangle + \langle x^k_B - x, u^k_B + \overline{C}x^k_B \rangle
\]
\[
\leq \frac{(\|z^* - x\| + (1 + \gamma/\beta)\|z^0 - z^*\| + \gamma\|Cf^*\|)\|z^0 - z^*\|}{\gamma\sqrt{\tau(k+1)}}\]

2. Suppose that $f$ and $\overline{A}$ are $L_f$ and $L_{\overline{A}}$-Lipschitz continuous respectively on the closed ball $B(x^*, (1 + \gamma/\beta)\|z^0 - z^*\|)$. Then the following convergence rate holds: For all $k \geq 0$ and $x \in \text{dom}(f) \cap \text{dom}(g)$, we have
\[
f(x^k_B) + g(x^k_B) + h(x^k_B) - (f + g + h)(x) + \langle x^k_B - x, \overline{A}x^k_B + u^k_B + \overline{C}x^k_B \rangle
\]
\[
\leq \frac{(\|z^* - x^*\| + (1 + \gamma/\beta)\|z^0 - z^*\| + \gamma\|Cf^*\|)\|z^0 - z^*\| + \gamma L_f \|z^0 - z^*\|}{\gamma\sqrt{\tau(k+1)}}
\]
\[
+ \frac{(1 + L_A)\|z^0 - z^*\|((1 + \gamma/\beta)(1 + L_A)\|z^0 - z^*\| + \|\overline{A}x^*\| + \|x^* - x\|)}{\sqrt{\tau(k+1)}}
\]

284
Thus, the convergence rates follow directly from Theorem C.1.

Proof. Fix $k \geq 0$.

Part 1: By Corollary C.2, we have

$$f(x_B^k) + g(x_B^k) + h(x_B^k) - (f + g + h)(x) + \langle x_A^k - x, u_B^k \rangle + \langle x_B^k - x, u_B^k + \overline{C}x_B^k \rangle$$

$$\leq \frac{1}{2\gamma} \kappa(k, x)$$

Thus, the convergence rates follow directly from Theorem C.1.

Part 2: Note that $f(x_B^k) - f(x_A^k) \leq L_f \|x_A^k - x_B^k\|$ by Lemma C.1. Because $x_B^k - x_A^k = z^k - Tz^k$, we have

$$f(x_B^k) + g(x_B^k) + h(x_B^k) - (f + g + h)(x)$$

$$\leq f(x_A^k) + g(x_B^k) + h(x_B^k) - (f + g + h)(x) + L_f \|x_A^k - x_B^k\|$$

$$\leq f(x_A^k) + g(x_B^k) + h(x_B^k) - (f + g + h)(x) + \frac{\|z^0 - z^*\|}{\sqrt{k+1}}.$$

Also,

$$\langle x_A^k - x, \overline{A}x_A^k \rangle$$

$$= \langle x_A^k - x_B^k, \overline{A}x_A^k \rangle + \langle x_B^k - x, \overline{A}x_A^k \rangle$$

$$= \langle x_A^k - x_B^k, \overline{A}x_A^k \rangle + \langle x_B^k - x, \overline{A}x_A^k - \overline{A}x_B^k \rangle + \langle x_B^k - x, \overline{A}x_B^k \rangle$$

$$\leq \|x_A^k - x_B^k\| \|\overline{A}x_A^k\| + \|x_B^k - x\| \|\overline{A}x_A^k - \overline{A}x_B^k\| + \langle x_B^k - x, \overline{A}x_B^k \rangle$$

$$\leq \frac{(1 + L_{\overline{A}})\|z^0 - z^*\| (\|\overline{A}x_A^k\| + \|\overline{A}x_B^k\|)}{\sqrt{k+1}} + \langle x_B^k - x, \overline{A}x_B^k \rangle$$

and for $x^* = J_{\gamma_B}(z^*)$,

$$\|\overline{A}x_A^k\| + \|x_B^k - x\| \leq \|\overline{A}x_A^k - \overline{A}x^*\| + \|\overline{A}x^*\| + \|x_B^k - x^*\| + \|x^* - x\|$$

$$\leq (1 + \gamma/\beta)(1 + L_{\overline{A}})\|z^0 - z^*\| + \|\overline{A}x^*\| + \|x^* - x\|$$

Thus, the rate follows by Part 1.
Corollary C.5 (Ergodic convergence rates of function values for Equation \((6.1.6)\)). Suppose that \((z^j)_{j \geq 0}\) is generated by Equation \((6.1.4)\), with \(A = \partial f, B = \partial g\) and \(C = \nabla h\). Let the assumptions be as in Theorem C.2. For all \(k \geq 0\), let \(x^f_k = (1/\sum_{i=0}^k \lambda_i) \sum_{i=0}^k \lambda_i x^f_i\), and let \(x^g_k = (1/\sum_{i=0}^k \lambda_i) \sum_{i=0}^k \lambda_i x^g_i\). Let \(x^* = J_{\gamma B}(z^*)\). Then the following convergence rates hold:

1. For all \(k \geq 0\), we have
   \[
   -2\|z^0 - z^*\| ||u_B^* + Cx^*||
   \sum_{i=0}^k \lambda_i 
   \leq f(x^f_k) + g(x^g_k) + h(x^g_k) - (f + g + h)(x^*) 
   \leq \frac{\|z^0 - x^*\|^2 + (\lambda_{2\beta - \gamma}) \|z^0 - z^*\|^2 + 4\gamma \|z^0 - z^*\||C x^*||}{2\gamma \sum_{i=0}^k \lambda_i}.
   \]

2. Suppose that \(f\) is \(L\)-Lipschitz continuous on \(B(x^*, (1 + \gamma/\beta)\|z^0 - z^*\|)\). Then the following convergence rate holds:
   \[
   0 \leq f(x^f_k) + g(x^g_k) + h(x^g_k) - (f + g + h)(x^*) 
   \leq \frac{\|z^0 - x^*\|^2 + (\lambda_{2\beta - \gamma}) \|z^0 - z^*\|^2 + 4\gamma \|z^0 - z^*\||C x^*|| + 4\gamma L \|z^0 - z^*\|}{2\gamma \sum_{i=0}^k \lambda_i}.
   \]

Proof. Fix \(k \geq 0\).

Part 1: We have the lower bound:

\[
\begin{align*}
&(f(x^f_k) + g(x^g_k) + h(x^g_k) - (f + g + h)(x^*)) \\
&\geq \langle x^f_k - x^*, \tilde{\nabla} f(x^*) \rangle + \langle x^g_k - x^*, \tilde{\nabla} g(x^*) + \nabla h(x^*) \rangle \\
&= \langle x^g_k - x^f_k, \tilde{\nabla} g(x^*) + \nabla h(x^*) \rangle.
\end{align*}
\]

where \(\tilde{\nabla} g(x^*) + \tilde{\nabla} f(x^*) + \nabla h(x^*) = 0\). In addition,

\[
\begin{align*}
&(f(x^f_k) + g(x^g_k) + h(x^g_k) - (f + g + h)(x^*)) \\
&\leq \frac{1}{2\gamma \sum_{i=0}^k \lambda_i} \sum_{i=0}^k \lambda_i k_{1i}^i(\lambda_i, x^*)
\end{align*}
\]

by Jensen’s inequality and Corollary C.1. Thus, the convergence rate follows by Theorem C.2.
Part 2: Note that \( f(\mathbf{x}_f^k) - f(\mathbf{x}_g^k) \leq L\|\mathbf{x}_f^k - \mathbf{x}_g^k\| \) by Lemma C.1 because the ball \( \bar{B}(x^*, (1 + \gamma/\beta)\|z^0 - z^*\|) \) is convex so the averaged sequences \((\mathbf{x}_f)_j \geq 0\) and \((\mathbf{x}_g)_j \geq 0\) must continue to lie in the ball. Therefore,

\[
f(\mathbf{x}_f^k) + g(\mathbf{x}_g^k) + h(\mathbf{x}_g^k) - (f + g + h)(x^*)
\leq f(\mathbf{x}_f^k) + g(\mathbf{x}_g^k) + h(\mathbf{x}_g^k) - (f + g + h)(x^*) + L\|\mathbf{x}_f^k - \mathbf{x}_g^k\|
\leq f(\mathbf{x}_f^k) + g(\mathbf{x}_g^k) + h(\mathbf{x}_g^k) - (f + g + h)(x^*) + \frac{2\|z^0 - z^*\|}{\sum_{i=0}^{k} \lambda_i}.
\]

Thus, the rate follows by Part 1. \qed

**Corollary C.6** (Ergodic convergence of variational inequalities for Equation (6.1.6)).

Suppose that \((z^i)_j \geq 0\) is generated by Equation (6.1.4), with \(A = \partial f + \bar{A}, B = \partial g + \bar{B}\) and \(C = \nabla h + \bar{C}\) as in Corollary C.2. In addition, suppose that \(\bar{A}\) and \(\bar{B}\) are skew linear maps (i.e., \(A^* = -A\), and \(B^* = -B\)), and suppose that \(\bar{C} \equiv 0\). Let the assumptions be as in Theorem C.2. For all \(k \geq 0\), let \(\mathbf{x}_A^k = (1/\sum_{i=0}^{k} \lambda_i) \sum_{i=0}^{k} \lambda_i x_i^k\), and let \(\mathbf{x}_B^k = (1/\sum_{i=0}^{k} \lambda_i) \sum_{i=0}^{k} \lambda_i x_i^k\). Then the following convergence rates hold:

1. For all \(k \geq 0\) and \(x \in \text{dom}(f) \cap \text{dom}(g)\), we have

\[
f(\mathbf{x}_A^k) + g(\mathbf{x}_B^k) + h(\mathbf{x}_B^k) - (f + g + h)(x) + \langle -x, \bar{A}\mathbf{x}_B^k + \bar{B}\mathbf{x}_B^k \rangle
\leq \frac{\|z^0 - x\|^2 + \gamma \|z^0 - z^*\|^2 + 4\gamma \|z^0 - z^*\|^2 \|C\mathbf{x}^*\| + 4\gamma \|\bar{A}\| \|x\| \|z^0 - z^*\|}{2\gamma \sum_{i=0}^{k} \lambda_i}
\]

2. Suppose that \(f\) is \(L_f\)-Lipschitz continuous on \(\bar{B}(x^*, (1 + \gamma/\beta)\|z^0 - z^*\|)\). Then the following convergence rate holds: For all \(k \geq 0\) and \(x \in \text{dom}(f) \cap \text{dom}(g)\), we have

\[
f(\mathbf{x}_B^k) + g(\mathbf{x}_B^k) + h(\mathbf{x}_B^k) - (f + g + h)(x) + \langle -x, \bar{A}\mathbf{x}_B^k + \bar{B}\mathbf{x}_B^k \rangle
\leq \frac{\|z^0 - x\|^2 + \gamma \|z^0 - z^*\|^2}{2\gamma \sum_{i=0}^{k} \lambda_i}
+ \frac{4\gamma \|z^0 - z^*\| \|C\mathbf{x}^*\| + 4\gamma \|\bar{A}\| \|x\| \|z^0 - z^*\| + 4\gamma L_f \|z^0 - z^*\|}{2\gamma \sum_{i=0}^{k} \lambda_i}.
\]

**Proof.** Fix \(k \geq 0\).
Part 1: By Corollary C.2, we have

\[ f(\bar{x}_A^k) + g(\bar{x}_B^k) + h(\bar{x}_B^k) - (f + g + h)(x) + \langle -x, A\bar{x}_B^k + B\bar{x}_B^k \rangle \]

\[ \leq \frac{1}{2\gamma} \sum_{i=0}^{k} \lambda_i \kappa_i^k(\lambda_i, x) + \langle x, \bar{A}(\bar{x}_B^k - \bar{x}_A^k) \rangle \]

\[ \leq \frac{1}{2\gamma} \sum_{i=0}^{k} \lambda_i \kappa_i^k(\lambda_i, x) + \frac{2\|A\|\|x\|\|z^0 - z^*\|}{\sum_{i=0}^{k} \lambda_i} \]

where we use the self orthogonality of skew symmetric maps (\(\langle Ay, y \rangle = \langle By, y \rangle = 0\) for all \(y \in H\)) and Jensen’s inequality. Thus, the convergence rates follow directly from Theorem C.2.

Part 2: Note that

\[ f(\bar{x}_B^k) - f(\bar{x}_A^k) \leq L_f \|\bar{x}_A^k - \bar{x}_B^k\| \] by Lemma C.1. Therefore,

\[ f(\bar{x}_B^k) + g(\bar{x}_B^k) + h(\bar{x}_B^k) - (f + g + h)(x) \]

\[ \leq f(\bar{x}_A^k) + g(\bar{x}_B^k) + h(\bar{x}_B^k) - (f + g + h)(x) + L_f \|x_A^k - x_B^k\| \]

\[ \leq f(\bar{x}_A^k) + g(\bar{x}_B^k) + h(\bar{x}_B^k) - (f + g + h)(x) + \frac{2L_f \|z^0 - z^*\|}{\sum_{i=0}^{k} \lambda_i}. \]

Thus, the rate follows by Part 1.

\[ \square \]

Corollary C.7 (Ergodic convergence rates of function values for Equation (6.1.7)). Suppose that \((z^j)_{j \geq 0}\) is generated by Equation (6.1.4), with \(A = \partial f, B = \partial g\) and \(C = \nabla h\). Let the assumptions be as in Theorem C.3. For all \(k \geq 0\), let \(\bar{x}_f^k = \frac{2}{((k + 1)(k + 2))} \sum_{i=0}^{k}(i + 1)x_f^i\), and let \(\bar{x}_g^k = \frac{2}{((k + 1)(k + 2))} \sum_{i=0}^{k}(i + 1)x_g^i\). Let \(x^* = J_B(z^*)\).

Then the following convergence rates hold:

1. For all \(k \geq 0\), we have

\[ \frac{-5\|z^0 - z^*\|}{\lambda(k + 1)} \]

\[ \leq f(\bar{x}_f^k) + g(\bar{x}_g^k) + h(\bar{x}_g^k) - (f + g + h)(x^*) \]

\[ \leq \frac{2\|z^* - x^*\|^2 + \left(2 + \frac{\gamma}{(2\beta - \gamma)}\right)\|z^0 - z^*\|^2 + 10\|z^0 - z^*\|\|Cx^*\|}{\gamma\lambda(k + 1)}. \]
2. Suppose that \( f \) is \( L \)-Lipschitz continuous on \( \overline{B(x^*, (1 + \gamma/\beta)\|z^0 - z^*\|)} \). Then the following convergence rate holds:

\[
0 \leq f(\bar{x}_f^k) + g(\bar{x}_g^k) + h(\bar{x}_g^k) - (f + g + h)(x^*) \leq \frac{2\|z^* - x^*\|^2 + \left(2 + \frac{\gamma}{(2\beta - \gamma)}\right)\|z^0 - z^*\|^2}{\gamma\lambda(k + 1)} + \frac{10\|z^0 - z^*\|\|Cx^*\| + 5\gamma L_f\|z^0 - z^*\|}{\gamma\lambda(k + 1)} \quad \text{(C.28)}
\]

**Proof.** Fix \( k \geq 0 \).

**Part 1:** We have the lower bound:

\[
f(\bar{x}_f^k) + g(\bar{x}_g^k) + h(\bar{x}_g^k) - (f + g + h)(x^*) \geq \langle \bar{x}_f^k - x^*, \nabla f(x^*) \rangle + \langle \bar{x}_g^k - x^*, \nabla g(x^*) + \nabla h(x^*) \rangle = \langle \bar{x}_g^k - \bar{x}_f^k, \nabla g(x^*) + \nabla h(x^*) \rangle.
\]

where \( \nabla g(x^*) + \nabla f(x^*) + \nabla h(x^*) = 0 \). In addition,

\[
f(\bar{x}_f^k) + g(\bar{x}_g^k) + h(\bar{x}_g^k) - (f + g + h)(x^*) \leq \frac{2}{2\gamma\lambda(k + 1)(k + 2)} \sum_{i=0}^{k} (i + 1)\kappa_i^1(\lambda, x^*)
\]

by Jensen’s inequality and Corollary C.1. Thus, the convergence rate follows by Theorem C.3.

**Part 2:** Note that \( f(\bar{x}_g^k) - f(\bar{x}_f^k) \leq L\|\bar{x}_f^k - \bar{x}_g^k\| \) by Lemma C.1 because the ball \( \overline{B(x^*, (1 + \gamma/\beta)\|z^0 - z^*\|)} \) is convex, so the averaged sequences \((\bar{x}_f)_{j \geq 0}\) and \((\bar{x}_g)_{j \geq 0}\) must continue to lie in the ball. Therefore,

\[
f(\bar{x}_g^k) + g(\bar{x}_g^k) + h(\bar{x}_g^k) - (f + g + h)(x^*) \leq f(\bar{x}_f^k) + g(\bar{x}_g^k) + h(\bar{x}_g^k) - (f + g + h)(x^*) + L\|\bar{x}_f^k - \bar{x}_g^k\| \leq f(\bar{x}_f^k) + g(\bar{x}_g^k) + h(\bar{x}_g^k) - (f + g + h)(x^*) + \frac{5L\|z^0 - z^*\|}{\lambda(k + 1)}.
\]

Thus, the rate follows by Part 1. \( \square \)
Corollary C.8 (Ergodic convergence of variational inequalities for Equation (6.1.7)). Suppose that $(z^i)_{i \geq 0}$ is generated by Equation (6.1.4), with $A = \partial f + \overline{A}, B = \partial g + \overline{B}$ and $C = \nabla h + \overline{C}$ as in Corollary C.2. In addition, suppose that $\overline{A}$ and $\overline{B}$ are skew linear maps (i.e., $A^* = -A$, and $B^* = -B$), and suppose that $\overline{C} \equiv 0$. Let the assumptions be as in Theorem C.3. For all $k \geq 0$, let $x_A^k = (2/((k+1)(k+2))) \sum_{i=0}^{k} (i+1)x_A^i$, and let $x_B^k = (2/((k+1)(k+2))) \sum_{i=0}^{k} (i+1)x_B^i$. Then the following convergence rates hold:

1. For all $k \geq 0$ and $x \in \text{dom}(f) \cap \text{dom}(g)$, we have

$$f(x_A^k) + g(x_B^k) + h(x_B^k) - (f + g + h)(x) + \langle -x, A x_A^k + B x_B^k \rangle$$

$$\leq 2 \frac{\|z^* - x\|^2 + \left(2 + \frac{\gamma}{\sqrt{3z^* - \gamma}}\right) \|z^0 - z^*\|^2}{\lambda(k+1)}$$

$$\quad + \frac{10 \|z^0 - z^*\|\|C x^*\| + 5 \gamma \|A\|\|x\|\|z^0 - z^*\|}{\lambda(k+1)} + 10 \|z^0 - z^*\|. \quad (C.33)$$

2. Suppose that $f$ is $L_f$-Lipschitz continuous on $\overline{B}(x^*, (1 + \gamma/\beta)\|z^0 - z^*\|)$. Then the following convergence rate holds: For all $k \geq 0$ and $x \in \text{dom}(f) \cap \text{dom}(g)$, we have

$$f(x_B^k) + g(x_B^k) + h(x_B^k) - (f + g + h)(x) + \langle -x, A x_B^k + B x_B^k \rangle$$

$$\leq 2 \frac{\|z^* - x\|^2 + \left(2 + \frac{\gamma}{\sqrt{3z^* - \gamma}}\right) \|z^0 - z^*\|^2}{\lambda(k+1)}$$

$$\quad + \frac{10 \|z^0 - z^*\|\|C x^*\| + 5 \gamma \|A\|\|x\|\|z^0 - z^*\| + 5 \gamma L_{\overline{f}} \|z^0 - z^*\|}{\lambda(k+1)} + 10 \|z^0 - z^*\|. \quad \text{Part 1:}$$

$$f(x_A^k) + g(x_B^k) + h(x_B^k) - (f + g + h)(x) + \langle -x, A x_B^k + B x_B^k \rangle$$

$$\leq \frac{2}{2 \lambda(k+1)(k+2)} \sum_{i=0}^{k} (i+1) \kappa_1^k(\lambda, x) + \langle x, A(x_B^k - x_A^k) \rangle$$

$$\quad \leq \frac{2}{2 \lambda(k+1)(k+2)} \sum_{i=0}^{k} (i+1) \kappa_1^k(\lambda, x) + \frac{5 \|A\|\|x\|\|z^0 - z^*\|}{\lambda(k+1)}. \quad (C.33)$$

Proof. Fix $k \geq 0$.

Part 1: By Corollary C.2, we have

$$f(x_A^k) + g(x_B^k) + h(x_B^k) - (f + g + h)(x) + \langle -x, A x_B^k + B x_B^k \rangle$$

$$\leq \frac{2}{2 \lambda(k+1)(k+2)} \sum_{i=0}^{k} (i+1) \kappa_1^k(\lambda, x) + \langle x, A(x_B^k - x_A^k) \rangle$$

$$\quad \leq \frac{2}{2 \lambda(k+1)(k+2)} \sum_{i=0}^{k} (i+1) \kappa_1^k(\lambda, x) + \frac{5 \|A\|\|x\|\|z^0 - z^*\|}{\lambda(k+1)}. \quad (C.33)$$

290
where we use the self orthogonality of skew symmetric maps \((\langle Ay, y \rangle = \langle By, y \rangle = 0\) for all \(y \in H\)) and Jensen’s inequality. Thus, the convergence rates follow directly from Theorem C.3.

Part 2: Note that \(f(x_k^B) - f(x_k^A) \leq L_f \|x_k^A - x_k^B\|\) by Lemma C.1. Therefore,
\[
\begin{align*}
&f(x_k^B) + g(x_k^B) + h(x_k^B) - (f + g + h)(x) \\
&\leq f(x_k^A) + g(x_k^B) + h(x_k^B) - (f + g + h)(x) + L_f \|x_k^A - x_k^B\| \quad \text{(C.33)}
\end{align*}
\]
Thus, the rate follows by Part 1.

C.4.3 Strong monotonicity

In this section, we deduce the convergence rates of the terms \(Q(\cdot, \cdot)\) under general assumptions.

Corollary C.9 (Strong convergence). Suppose that \((z^i)_{j \geq 0}\) is generated by Equation (6.1.4). Let \(z^*\) be a fixed point of \(T\) and let \(x^* = J_{\gamma B}(z^*)\). Then for all \(k \geq 0\), the following convergence rates hold:

1. **Nonergodic convergence:** Let the assumptions of Theorem C.1 hold. Then
\[
Q_A(x_k^A, x^*) + Q_B(x_k^B, x^*) + Q_C(x_k^C, x^*) \leq \frac{(1 + \gamma/\beta) \|z^0 - z^*\|^2}{\gamma \sqrt{T}(k+1)}
\]
and \(Q_A(x_k^A, x^*) + Q_B(x_k^B, x^*) + Q_C(x_k^C, x^*) = o\left(\frac{1}{\sqrt{k+1}}\right)\).

2. **“Best” iterate convergence:** Let the assumptions of Theorem C.1 hold. Suppose that \(\Lambda := \inf_{j \geq 0} \lambda_j\). Then
\[
\min_{i=0, \ldots, k} \left\{Q_A(x_i^A, x^*) + Q_B(x_i^B, x^*) + Q_C(x_i^C, x^*)\right\} \leq \frac{\left(1 + \frac{\gamma}{2\gamma - \gamma}\right) \|z^0 - z^*\|^2}{2 \gamma \Lambda (k+1)}
\]
and \(\min_{i=0, \ldots, k} \left\{Q_A(x_i^A, x^*) + Q_B(x_i^B, x^*) + Q_C(x_i^C, x^*)\right\} = o\left(\frac{1}{(k+1)}\right)\).
3. **Ergodic convergence for Equation (6.1.6):** Let the assumptions for Theorem C.2 hold. Then

\[
\frac{1}{\sum_{i=0}^{k} \lambda_i} \sum_{i=0}^{k} \lambda_i \left( Q_A(x_A^k, x^*) + Q_B(x_B^k, x^*) + Q_C(x_B^k, x^*) \right) \\
\leq \left( 1 + \frac{\gamma}{2\beta \varepsilon - \gamma} \right) \frac{\|z^0 - z^*\|^2}{2\gamma \sum_{i=0}^{k} \lambda_i}.
\]

4. **Ergodic convergence for Equation (6.1.7):** Let the assumptions for Theorem C.3 hold. Then

\[
\frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} (i+1) \left( Q_A(x_A^k, x^*) + Q_B(x_B^k, x^*) + Q_C(x_C^k, x^*) \right) \\
\leq \left( 1 + \frac{\gamma}{2\beta \varepsilon - \gamma} \right) \frac{\|z^0 - z^*\|^2}{\gamma \lambda (k+1)}.
\]

**Proof.** The “best” iterate convergence result follows Lemma 2.2.1 of Chapter 2 because we have

\[
\sum_{i=0}^{\infty} 2\gamma \Delta(Q_A(x_A^i, x^*) + Q_B(x_B^i, x^*) + Q_C(x_C^i, x^*)) \leq \sum_{i=0}^{\infty} 2\gamma \lambda_i (Q_A(x_A^i, x^*) + Q_B(x_B^i, x^*) + Q_C(x_C^i, x^*)) \leq \sum_{i=0}^{\infty} \lambda_i k_i^2 (\lambda_i, x^*) \leq (1 + \gamma/(2\beta \varepsilon - \gamma))
\]

by the upper bounds in Equations (C.16) and (C.26).

The rest of the results follow by combining the upper bound in Equation (C.16) with the convergence rates in Theorems C.1, C.2, and C.3. \(\square\)

At first glance it may be seem that the ergodic bounds in Theorem C.9 are not meaningful. However, whenever \(\mu_A > 0\), we can apply Jensen’s inequality to show that

\[
\sum_{i=0}^{k} \nu_i Q_A(x_A^i, x^*) \geq \mu_A \left\| \sum_{i=0}^{k} \nu_i x_A^i - x^* \right\|^2
\]

for any positive sequence of stepsizes \((\nu_j)^{k}_{j=0}\), such that \(\sum_{i=0}^{k} \nu_i = 1\). Thus, the ergodic bounds really prove strong convergence rates for the ergodic iterates generated by Equations (6.1.6) and (6.1.7).
C.4.4 Lipschitz differentiability

In this section, we focus on function minimization. In particular, we let \( A = \partial f \), \( B = \partial g \), and \( C = \nabla h \), where \( f, g \) and \( h \) are closed, proper, and convex, and \( \nabla h \) is \((1/\beta_f)\)-Lipschitz.

We make the following assumption regarding the regularity of \( f \):

The gradient of at least one of \( f \) is Lipschitz.

Under this assumption we will show that the “best” objective error after \( k \) iterations of Equation (6.1.4) has order \( o(1/(k + 1)) \).

The techniques of this section can also be applied to show a similar result for \( g \). The proof is somewhat more technical, so we omit it.

The following theorem will be used several times throughout our analysis. See [11, Theorem 18.15(iii)] for a proof.

**Theorem C.4** (Descent theorem). Suppose that \( f : \mathcal{H} \rightarrow (-\infty, \infty] \) is closed, convex, and differentiable. If \( \nabla f \) is \((1/\beta_f)\)-Lipschitz, then for all \( x, y \in \text{dom}(f) \), we have the upper bound

\[
f(x) \leq f(y) + \langle x - y, \nabla f(y) \rangle + \frac{1}{2\beta} \| x - y \|^2.
\]

(C.35)

**Proposition C.4** (Lipschitz differentiable upper bound). Suppose that \((z^j)_{j \geq 0}\) is generated by Equation (6.1.4). Then the following bounds hold: Suppose that \( f \) is differentiable and \( \nabla f \) is \((1/\beta_f)\)-Lipschitz. Then

\[
2\gamma \lambda ((f + g + h)(x_g) - (f + g + h)(x^*)) \leq \begin{cases} 
\| z - z^* \|^2 - \| z^+ - z^* \|^2 + \left( 1 + \frac{\gamma - \beta_f}{\beta_f \lambda} \right) \| z - z^+ \|^2 \\
2\gamma \langle \nabla h(x_g) - \nabla h(x^*), z - z^+ \rangle & \text{if } \gamma \leq \beta_f \\
\left( 1 + \frac{\gamma - \beta_f}{2\beta_f} \right) (\| z - z^* \|^2 - \| z^+ - z^* \|^2 + \| z - z^+ \|^2) \\
+ 2\gamma \left( 1 + \frac{\gamma - \beta_f}{2\beta_f} \right) \langle \nabla h(x_g) - \nabla h(x^*), z - z^+ \rangle & \text{if } \gamma > \beta_f.
\end{cases}
\]

(C.36)
Proof. Because $\nabla f$ is $(1/\beta_f)$-Lipschitz, we have

$$f(x_g) \leq f(x_f) + \langle x_g - x_f, \nabla f(x_f) \rangle + \frac{1}{2\beta_f} \|x_g - x_f\|^2; \quad (C.37)$$

$$S_f(x_f, x^*) \geq \frac{\beta_f}{2} \|\nabla f(x_f) - \nabla f(x^*)\|^2. \quad (C.38)$$

By applying the identity $z^* - x^* = \gamma \nabla g(x^*) = -\gamma \nabla f(x^*) - \gamma \nabla h(x^*)$, the cosine rule (6.1.12), and the identity $z - z^+ = \lambda(x_g - x_f)$ (see Lemma 6.3.1) multiple times, we have

$$2\langle z - z^+, z^* - x^* \rangle + 2\gamma \lambda \langle x_g - x_f, \nabla f(x_f) \rangle$$

$$= 2\lambda(x_g - x_f, \gamma \nabla g(x^*) + \gamma \nabla f(x_f))$$

$$= 2\lambda \langle \gamma \nabla g(x_g) + \gamma \nabla h(x_g) + \gamma \nabla f(x_f), \gamma \nabla f(x_f) - \gamma \nabla f(x^*) \rangle - 2\langle z - z^+, \gamma \nabla h(x^*) \rangle$$

$$= \lambda \left( \|\gamma \nabla f(x_f) - \gamma \nabla f(x^*)\|^2 + \|x_g - x_f\|^2 \right.$$ 

$$- \|\gamma \nabla g(x_g) + \gamma \nabla h(x_g) - \gamma \nabla g(x^*) - \gamma \nabla h(x^*)\|^2 \left. \right) - 2\langle z - z^+, \gamma \nabla h(x^*) \rangle. \quad (C.39)$$

By Lemma 6.3.1 (i.e., $z - z^+ = \lambda(x_g - x_f)$), we have

$$\left( 1 - \frac{2}{\lambda} \right) \|z - z^+\|^2 + \lambda \left( \frac{\gamma}{\beta_f} + 1 \right) \|x_g - x_f\|^2 = \left( 1 + \frac{(\gamma - \beta_f)}{\beta_f \lambda} \right) \|z - z^+\|^2.$$ 

Therefore,

$$2\gamma \lambda ((f + g + h)(x_g) - (f + g + h)(x^*))$$

$$(C.37) \leq 2\gamma \lambda (f(x_f) + g(x_g) + h(x_g) - (f + g + h)(x^*)) + 2\gamma \lambda \langle x_g - x_f, \nabla f(x_f) \rangle$$

$$+ \frac{\gamma \lambda}{\beta_f} \|x_g - x_f\|^2$$

$$(C.19) \leq \|z - z^*\|^2 - \|z^+ - z^*\|^2 + 2\langle z - z^+, z^* - x^* \rangle + 2\gamma \lambda \langle x_g - x_f, \nabla f(x_f) \rangle$$

$$+ \left( 1 - \frac{2}{\lambda} \right) \|z^+ - z\|^2 + 2\gamma \langle \nabla h(x_g), z - z^+ \rangle + \frac{\gamma \lambda}{\beta_f} \|x_g - x_f\|^2 - 2\gamma \lambda S_f(x_f, x^*)$$

$$(C.39) \leq \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left( 1 - \frac{2}{\lambda} \right) \|z - z^+\|^2 + \lambda \left( \frac{\gamma}{\beta_f} + 1 \right) \|x_g - x_f\|^2$$

$$+ \lambda \|\gamma \nabla f(x_f) - \gamma \nabla f(x^*)\|^2 + 2\gamma \langle \nabla h(x_g) - \nabla h(x^*), z - z^+ \rangle - 2\gamma \lambda S_f(x_f, x^*)$$

$$\leq \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left( 1 + \frac{(\gamma - \beta_f)}{\beta_f \lambda} \right) \|z - z^+\|^2$$

$$(C.38) + 2\gamma \langle \nabla h(x_g) - \nabla h(x^*), z - z^+ \rangle + \gamma \lambda (\gamma - \beta_f) \|\nabla f(x_f) - \nabla f(x^*)\|^2.$$ 

(C.40)
If $\gamma \leq \beta f$, then we can drop the last term. If $\gamma > \beta f$, then we apply the upper bound in Equation (C.20) to get:

$$
\gamma \lambda (\gamma - \beta f) \| \nabla f(x_f) - \nabla f(x^*) \|^2 \\
\leq \frac{\gamma - \beta f}{2\beta f} \left( \| z - z^* \|^2 - \| z^+ - z^* \|^2 + \left( 1 - \frac{2}{\lambda} \right) \| z^+ - z \|^2 \right) \\
+ 2\gamma \langle \nabla h(x_g) - \nabla h(x^*), z - z^+ \rangle.
$$

The result follows by using the above inequality in Equation (C.40) together with the following identity:

$$
\left( 1 + \frac{\gamma - \beta f}{\beta f \lambda} \right) \| z - z^+ \|^2 + \frac{\gamma - \beta f}{2\beta f} \left( 1 - \frac{2}{\lambda} \right) \| z - z^+ \|^2 = \left( 1 + \frac{\gamma - \beta f}{2\beta f} \right) \| z - z^+ \|^2.
$$

\[ \square \]

**Theorem C.5** ("Best" objective error rate). Let $(z^j)_{j \geq 0}$ be generated by Equation (6.1.4) with $\gamma \in (0, 2\beta)$ and $\tau = \inf_{j \geq 0} \lambda_j (1 - \alpha \lambda_j) / \alpha > 0$. Then the following bound holds: If $f$ is differentiable and $\nabla f$ is $(1/\beta f)$-Lipschitz, then

$$
0 \leq \min_{i=0,\ldots,k} \left\{ (f + g + h)(x^i) - (f + g + h)(x^*) \right\} = o \left( \frac{1}{k+1} \right).
$$

**Proof.** By Lemma 2.2.1 of Chapter 2, it suffices to show that all of the upper bounds in Proposition C.4 are summable. In both of the cases, the alternating sequence (and any constant multiple) $(\| z^j - z^* \|^2 - \| z^{j+1} - z^* \|^2)_{j \geq 0}$ is clearly summable. In addition, we know that $(\| z^j - z^{j+1} \|^2)_{j \geq 0}$ is summable by Part 1 of Theorem 6.3.1, and every coefficient of this sequence in the two upper bounds is bounded (because $(\lambda_j)_{j \geq 0}$ is a bounded sequence). Thus, the part pertaining to $(\| z^j - z^{j+1} \|^2)_{j \geq 0}$ is summable.

Finally, we just need to show that $(\langle \nabla h(x^j_g) - \nabla h(x^*), z^j - z^{j+1} \rangle)_{j \geq 0}$ is summable. The Cauchy-Schwarz inequality and Young’s inequality for real numbers show that for all $k \geq 0$, we have

$$
2\langle \nabla h(x^j_g) - \nabla h(x^*), z^k - z^{k+1} \rangle \leq \| \nabla h(x^j_g) - \nabla h(x^*) \|^2 + \| z^k - z^{k+1} \|^2.
$$

The second term is summable by the argument above, and the first term is summable by Part 4 of Theorem 6.3.1. \[ \square \]
Remark C.1. The order of convergence in Theorem C.5 is sharp by Theorem 2.7.3 of Chapter 2 and generalizes similar results the convergence rates presented in this thesis for Douglas-Rachford splitting, forward-backward splitting and forward-Douglas-Rachford splitting.

C.4.5 Linear convergence

In this section we show that Equation (6.1.4) converges linearly whenever \((\mu_A + \mu_B + \mu_C)(1/L_A + 1/L_B) > 0\)

where \(L_A\) and \(L_B\) are the Lipschitz constants of \(A\) and \(B\) and we follow the convention that \(1/L_A = 0\) or \(1/L_B = 0\) whenever \(A\) or \(B\) fail to be Lipschitz, respectively.

The first result of this section is an inequality that will help us deduce contraction factors for \(T\) in Theorem C.6.

Proposition C.5. Assume the setting of Theorem 6.3.1. In particular, let \(\varepsilon \in (0, 1)\), let \(\gamma \in (0, 2\beta\varepsilon)\), let \(\alpha = 1/(2-\varepsilon)\), and let \(\lambda \in (0, 1/\alpha)\). Let \(z \in \mathcal{H}\) and let \(z^+ = (1-\lambda)z + \lambda Tz\).

Let \(z^*\) be a fixed point of \(T\) and let \(x^* = J_{\gamma B}(z^*)\). Let \(x_A\) and \(x_B\) be defined as in Lemma 6.3.1. Let \(Q_A, Q_B\) and \(Q_C\) be defined as in Proposition C.2. Then the following
inequality holds:
\[
\|z^+ - z^*\|^2 + \left(1 - \frac{1}{\lambda \alpha}\right) \|z - z^+\|^2 + 2\gamma \lambda Q_A(x_A, x^*) + 2\gamma \lambda Q_B(x_B, x^*) + 2\gamma \lambda Q_C(x_B, x^*)
\]
\[
- \frac{\gamma^2 \lambda}{\varepsilon} \|C x_B - C x^*\|^2
\]
\[
\leq \|z - z^*\|^2
\]
\[
\leq \min \left\{ (1 + \gamma L_B)^2 \|x_B - x^*\|^2, \\
3 \left( (1 + \gamma L_A)^2 \|x_A - x^*\|^2 + \gamma^2 \|C x_B - C x^*\|^2 + 4\|x_B - x_A\|^2 \right), \\
3(1 + 2\gamma^2 L_B^2) (\|x_A - x^*\|^2 + \|x_A - x_B\|^2), \\
4 \left( (1 + 2\gamma^2 L_A^2) \|x_B - x^*\|^2 + \gamma^2 \|C x_B - C x^*\|^2 + (1 + 2\gamma^2 L_A^2) \|x_B - x_A\|^2 \right) \right\}
\]
(C.41)

Proof. Equation (C.41) shows that:
\[
\|z^+ - z^*\|^2 + \left(1 - \frac{1}{\lambda \alpha}\right) \|z - z^+\|^2 + 2\gamma \lambda Q_A(x_A, x^*) + 2\gamma \lambda Q_B(x_B, x^*) + 2\gamma \lambda Q_C(x_B, x^*)
\]
\[
\leq \|z - z^*\|^2 + 2\gamma \langle z - z^+, C x_B - C x^* \rangle.
\]

From Cauchy-Schwarz and Young’s inequality, we have
\[
2\gamma \langle z - z^+, C x_B - C x^* \rangle \leq \frac{\varepsilon}{\lambda} \|z - z^+\|^2 + \frac{\gamma^2 \lambda}{\varepsilon} \|C x_B - C x^*\|^2.
\]

The lower bound now follows by rearranging.

The upper bound follows from the following bounds (where we take $L_B = \infty$ or
\( L_A = \infty \) respectively whenever \( A \) or \( B \) fail to be Lipschitz):

\[
\|z - z^*\|^2 = \|x_B + \gamma u_B - (x^* + \gamma u_B^*)\|^2 \leq (1 + \gamma L_B)^2 \|x_B - x^*\|^2;
\]

\[
\|z - z^*\|^2 = \|x_B + \gamma u_B - (x^* - \gamma u_A^* - \gamma C x^*)\|^2
= \|x_B - \gamma(u_A + C x_B) + \gamma(u_B + u_A + C x_B) - (x^* - \gamma u_A^* - \gamma C x^*)\|^2
\leq \|x_A - \gamma(u_A + C x_B) + 2(x_B - x_A) - (x^* - \gamma u_A^* - \gamma C x^*)\|^2
\leq 3 \left( \|x_A - \gamma u_A - (x^* - \gamma u_A^*)\|^2 + \gamma^2 \|C x_B - C x^*\|^2 + 4\|x_B - x_A\|^2 \right)
\leq 3 \left( (1 + \gamma L_A)^2 \|x_A - x^*\|^2 + \gamma^2 \|C x_B - C x^*\|^2 + 4\|x_B - x_A\|^2 \right);
\]

\[
\|z - z^*\|^2 = \|x_B + \gamma u_B - (x^* + \gamma u_B^*)\|^2
= \|x_A + \gamma u_B - (x^* + \gamma u_B^*) + (x_B - x_A)\|^2
\leq 3 \left( \|x_A - x^*\|^2 + \gamma^2 \|u_B - u_B^*\|^2 + \|x_A - x_B\|^2 \right)
\leq 3 \left( \|x_A - x^*\|^2 + \gamma^2 L_B^2 \|x_B - x^*\|^2 + \|x_A - x_B\|^2 \right)
\leq 3(1 + 2\gamma^2 L_B^2) \left( \|x_A - x^*\|^2 + \|x_A - x_B\|^2 \right);
\]

\[
\|z - z^*\|^2 = \|x_B + \gamma u_B - (x^* + \gamma u_B^*)\|^2
= \|x_B - \gamma(u_A + C x_B) - (x^* - \gamma(u_A + C x^*)) + (x_B - x_A)\|^2
\leq 4 \left( \|x_B - x^*\|^2 + \gamma^2 \|u_A - u_A^*\|^2 + \gamma^2 \|C x_B - C x^*\|^2 + \|x_B - x_A\|^2 \right)
\leq 4 \left( \|x_B - x^*\|^2 + \gamma^2 L_A^2 \|x_A - x^*\|^2 + \|x_B - x_A\|^2 \right)
\leq 4 \left( (1 + 2\gamma^2 L_A^2) \|x_B - x^*\|^2 + \gamma^2 \|C x_B - C x^*\|^2 + (1 + 2\gamma^2 L_A^2) \|x_B - x_A\|^2 \right).
\]

The following theorem proves linear convergence of Equation (6.1.4) whenever \((\mu_A + \mu_B + \mu_C)(1/L_A + 1/L_B) > 0\).

**Theorem C.6.** Assume the setting of Theorem 6.3.1. In particular, let \( \varepsilon \in (0, 1) \), let \( \gamma \in (0, 2\beta \varepsilon) \), let \( \alpha = 1/(2 - \varepsilon) \), and let \( \lambda \in (0, 1/\alpha) \). Let \( z \in H \) and let \( z^+ = (1 - \lambda)z + \lambda Tz \). Let \( z^* \) be a fixed point of \( T \) and let \( x^* = J_{\gamma B}(z^*) \). Then the following inequality holds under each of the conditions below:

\[
\|z^+ - z^*\| \leq (1 - C(\lambda))^{1/2} \|z - z^*\|
\]

298
where $C(\lambda) \in [0, 1]$ is defined below under different scenarios.

1. Suppose that $B$ is $L_B$-Lipschitz, and $\mu_B$ strongly monotone. Then

$$C(\lambda) = \frac{2L_B \gamma \lambda}{(1 + \gamma L_B)^2}.$$  

2. Suppose that $A$ is $L_A$-Lipschitz and $\mu_A$-strongly monotone. Then

$$C(\lambda) = \frac{\lambda}{3} \min \left\{ \frac{2\mu_A \gamma}{(1 + \gamma L_A)^2}, \frac{\lambda}{4} \left( \frac{1}{\alpha \lambda} - 1 \right), \frac{2\beta - \gamma/\varepsilon}{\gamma} \right\}.$$  

3. Suppose that $A$ is $\mu_A$ strongly monotone and $B$ is $L_B$-Lipschitz. Then

$$C(\lambda) = \frac{\lambda}{3(1 + 2\gamma^2 L_B^2)} \min \left\{ 2\gamma \mu_A, \lambda \left( \frac{1}{\alpha \lambda} - 1 \right) \right\}.$$  

4. Suppose that $A$ is $L_A$-Lipschitz and $B$ is $\mu_B$-strongly monotone. Then

$$C(\lambda) = \frac{\lambda}{4} \min \left\{ \frac{2\gamma \mu_B}{(1 + 2\gamma^2 L_A^2)}, \frac{2\beta - \gamma/\varepsilon}{\gamma}, \frac{\lambda}{(1 + 2\gamma^2 L_A^2)} \left( \frac{1}{\alpha \lambda} - 1 \right) \right\}.$$  

5. Suppose that $A$ is $L_A$-Lipschitz and $C$ is $\mu_C$-strongly monotone. Let $\eta \in (0, 1)$ be large enough that $2\eta \beta > \gamma/\varepsilon$. Then

$$C(\lambda) = \frac{\lambda}{4} \min \left\{ \frac{2\gamma \mu_C (1 - \eta)}{(1 + 2\gamma^2 L_A^2)}, \frac{2\beta - \gamma/\varepsilon}{\gamma}, \frac{\lambda}{(1 + 2\gamma^2 L_A^2)} \left( \frac{1}{\alpha \lambda} - 1 \right) \right\}.$$  

6. Suppose that $B$ is $L_B$-Lipschitz and $C$ is $\mu_C$-strongly monotone. Let $\eta \in (0, 1)$ be large enough that $2\eta \beta > \gamma/\varepsilon$. Then

$$C(\lambda) = \frac{2\gamma \mu_C (1 - \eta)}{(1 + \gamma L_B)^2}.$$  

Proof. Each part of the proof is based on the following idea: If

$$a_0, \ldots, a_n, b_0, \ldots, b_n, c_0, \ldots, c_n \in \mathbb{R}_{++}$$

for some $n \geq 0$, and

$$\|z^+ - z^*\|^2 + \sum_{i=0}^n a_i c_i \leq \|z - z^*\|^2 \leq \sum_{i=0}^n a_i b_i,$$
then $\sum_{i=0}^{n} a_i b_i \leq \max\{b_i/c_i \mid i = 0, \cdots, n\} \sum_{i=0}^{n} a_i c_i$, so

$$\|z^+ - z^*\|^2 + \min\{c_i/b_i \mid i = 0, \cdots, n\}\|z - z^*\|^2 \leq \|z^+ - z^*\|^2 + \sum_{i=0}^{n} a_i c_i$$

$$\leq \|z - z^*\|^2 \leq \|z - z^*\|^2.$$ Thus, $$\|z^+ - z^*\| \leq (1 - \min\{c_i/b_i \mid i = 0, \cdots, n\})^{1/2} \|z - z^*\|.$$ In each case the terms $a_i c_i$ will be taken from the left hand side of Equation (C.41), and the terms $a_i b_i$ will be taken from the right of the same equation.

**Part 1:** We use the first upper bound in Equation (C.41) and set $a_0 = \|x_B - x^*\|^2, c_0 = 2\gamma \lambda \mu_A$, and $b_0 = (1 + \gamma L_B)^2$.

**Part 2:** We use the second upper bound in Equation (C.41) and set $a_0 = \|x_A - x^*\|^2, c_0 = 2\gamma \lambda \mu_A, b_0 = 3(1 + \gamma L_A)^2, a_1 = \|C x_B - C x^*\|^2, c_1 = \gamma \lambda (2\beta - \gamma/\varepsilon), b_1 = 3\gamma^2, a_2 = \|x_B - x_A\|, c_2 = \lambda^2(1/(\lambda \alpha) - 1)$, and $b_2 = 12$.

**Part 3:** We use the third upper bound in Equation (C.41) and set $a_0 = \|x_A - x^*\|^2, c_0 = 2\gamma \lambda \mu_A, b_0 = 3(1 + \gamma L_A)^2, a_1 = \|x_A - x_B\|^2, c_1 = \lambda^2(1/(\lambda \alpha) - 1)$, and $b_1 = 3(1 + 2\gamma^2 L_B^2)$.

**Part 4:** We use the fourth upper bound in Equation (C.41) and set $a_0 = \|x_B - x^*\|^2, c_0 = 2\gamma \lambda \mu_B, b_0 = 4(1 + 2\gamma^2 L_A^2), a_1 = \|C x_B - C x^*\|^2, c_1 = \gamma \lambda (2\beta - \gamma/\varepsilon), b_1 = 4\gamma^2, a_2 = \|x_B - x_A\|^2, c_2 = \lambda^2(1/(\lambda \alpha) - 1)$, and $b_2 = 4(1 + 2\gamma^2 L_A^2)$.

**Part 5:** We use the fourth upper bound in Equation (C.41) and set $a_0 = \|x_B - x^*\|^2, c_0 = 2\gamma \lambda \mu_C(1 - \eta), b_0 = 4(1 + 2\gamma^2 L_A^2), a_1 = \|C x_B - C x^*\|^2, c_1 = \gamma \lambda (2\eta \beta - \gamma/\varepsilon), b_1 = 4\gamma^2, a_2 = \|x_B - x_A\|^2, c_2 = \lambda^2(1/(\lambda \alpha) - 1)$, and $b_2 = 4(1 + 2\gamma^2 L_A^2)$.

**Part 6:** We use the first upper bound in Equation (C.41) and set $a_0 = \|x_B - x^*\|^2, c_0 = 2\gamma \lambda \mu_C(1 - \eta)$, and $b_0 = (1 + \gamma L_B)^2$.

**Remark C.2.** Note that the contraction factors can be improved whenever A or B are known to be subdifferential operators of convex functions because the function $Q(\cdot, \cdot)$ can be made larger with Proposition C.2. We do not pursue this here due to lack of space.
Remark C.3. Note that we can relax the conditions of Theorem C.6. Indeed, we only need to assume that $C$ is Lipschitz to derive linear convergence, not necessarily cocoercive. We do not pursue this extension here due to lack of space.

C.4.6 Arbitrarily slow convergence when $\mu_C \mu_A > 0$.

This section shows that the result of Theorem C.6 cannot be improved in the sense that we cannot expect linear convergence even if $C$ and $A$ are strongly monotone. The results of this section parallel similar results shown in Chapter 5.

The main example

Let $\mathcal{H} = \ell_2^2(\mathbb{N}) = \mathbb{R}^2 \oplus \mathbb{R}^2 \oplus \cdots$. Let $R_\theta$ denote counterclockwise rotation in $\mathbb{R}^2$ by $\theta$ degrees. Let $e_0 := (1,0)$ denote the standard unit vector, and let $e_\theta := R_\theta e_0$. Suppose that $(\theta_j)_{j \geq 0}$ is a sequence of angles in $(0, \pi/2]$ such that $\theta_i \to 0$ as $i \to \infty$. For all $i \geq 0$, let $c_i := \cos(\theta_i)$. We let

$$V := \mathbb{R}^2 e_0 \oplus \mathbb{R}^2 e_0 \oplus \cdots \quad \text{and} \quad U := \mathbb{R}^2 e_{\theta_0} \oplus \mathbb{R}^2 e_{\theta_1} \oplus \cdots.$$  \hspace{1cm} (C.42)

Note that [6, Section 7] proves the projection identities

$$ (P_U)_i = \begin{bmatrix} \cos^2(\theta_i) & \sin(\theta_i) \cos(\theta_i) \\ \sin(\theta_i) \cos(\theta_i) & \sin^2(\theta_i) \end{bmatrix} \quad \text{and} \quad (P_V)_i = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. $$

We now begin our extension of this example. Choose $a \geq 0$ and set $f = \iota_U + (a/2) \| \cdot \|^2$, $g = \iota_V$, and $h = (1/2) \| \cdot \|^2$. Set $A = \partial f$, $B = \partial g$, and $C = \nabla h$. Note that $\mu_h = 1$ and $\mu_f = a$. Thus, $\nabla h$ is 1-Lipschitz, and, hence, $\beta = 1$ and we can choose $\gamma = 1 < 2\beta$. Therefore, $\alpha = 2\beta/(4\beta - \gamma) = 2/3$, so we can choose $\lambda_k \equiv 1 < 1/\alpha$. We also note that $\text{prox}_{\gamma f} = (1/(1 + a)) P_V$. 

301
For all $i \geq 0$, we have

$$T_i := \frac{1}{a+1} (P_U)_i (2(P_V)_i - I_{R^2} - I_{R^2}) + I_{R^2} - (P_V)_i$$

$$= \frac{1}{a+1} (P_U)_i \begin{bmatrix} 0 & 0 \\ 0 & -2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \frac{1}{a+1} \begin{bmatrix} 0 & -2 \sin(\theta_i) \cos(\theta_i) \\ 0 & -2 \sin^2(\theta_i) + a + 1 \end{bmatrix}$$

where $T = \bigoplus_{i=0}^{\infty} T_i$ is the operator defined in Equation (6.1.3). Note that for all $i \geq 0$, the operator $(T)_i$ has eigenvector

$$z_i = \left( -\frac{2 \cos(\theta_i) \sin(\theta_i)}{1 + a - 2 \sin^2(\theta_i)}, 1 \right)$$

with eigenvalue $b_i := (a - 2(1 - c_i)^2 + 1)/(a + 1)$. Each component also has the eigenvector $(1, 0)$ with eigenvalue 0. Thus, the only fixed point of $T$ is $0 \in \mathcal{H}$. Finally, we note that

$$\|z_i\|^2 = \frac{4c_i^2(1 - c_i^2)}{(1 + a - 2(1 - c_i)^2)^2} + 1. \quad (C.43)$$

**Slow convergence proofs**

Part 2 of Theorem 6.3.1 shows that $z^{k+1} - z^k \to 0$. The following result is a consequence of [11, Proposition 5.27].

**Lemma C.2** (Strong convergence). *Any sequence $(z^j)_{j \geq 0} \subseteq \mathcal{H}$ generated by Algorithm 6.1.1 converges strongly to 0.*

The next Lemma appeared in Chapter 2

**Lemma C.3** (Arbitrarily slow sequence convergence). *Suppose that $F : \mathbf{R}_+ \to (0, 1)$ is a function that is monotonically decreasing to zero. Then there exists a monotonic sequence $(b_j)_{j \geq 0} \subseteq (0, 1)$ such that $b_k \to 1^-$ as $k \to \infty$ and an increasing sequence of integers $(n_j)_{j \geq 0} \subseteq \mathbf{N} \cup \{0\}$ such that for all $k \geq 0$,

$$\frac{b_{n_k}^{k+1}}{n_k + 1} > F(k + 1)e^{-1}. \quad (C.44)$$
The following is a simple corollary of Lemma C.3; the lemma first appeared in Corollary 5.6.1 of Chapter 5.

**Corollary C.10.** Let the notation be as in Lemma C.3. Then for all \( \eta \in (0, 1) \), we can find a sequence \((b_j)_{j \geq 0} \subseteq (\eta, 1)\) that satisfies the conditions of the lemma.

We are now ready to show that FDRS can converge arbitrarily slowly.

**Theorem C.7** (Arbitrarily slow convergence of (6.1.4)). For every function \( F : \mathbb{R}_{+} \rightarrow (0, 1) \) that strictly decreases to zero, there is a point \( z^0 \in \ell^2_{2}(N) \) and two closed subspaces \( U \) and \( V \) with zero intersection, \( U \cap V = \{0\} \), such that sequence \((z^j)_{j \geq 0}\) generated by Equation (6.1.4) applied to the functions \( f = \chi_U + (a/2)\|\cdot\|^2 \) and \( g = (1/2)\|\cdot\|^2 \), relaxation parameters \( \lambda_k \equiv 1 \), and stepsize \( \gamma = 1 \) satisfies the following bound:

\[
\|z^k - z^*\| \geq e^{-1}F(k),
\]

but \((\|z^j - z^*\|)_{j \geq 0}\) converges to 0.

**Proof.** For all \( i \geq 0 \), define \( z^0_i = (1/\|z_i\|(i + 1))z_i \), then \( \|z^0_i\| = 1/(i + 1) \) and \( z^0 \) is an eigenvector of \((T)_i\) with eigenvalue \( b_i = (a-2(1-c_i)^2+1)/(a+1) \). Define the concatenated vector \( z^0 = (z^0_i)_{i \geq 0} \). Note that \( z^0 \in \mathcal{H} \) because \( \|z^0\|^2 = \sum_{i=0}^{\infty} 1/(i + 1)^2 < \infty \). Thus, for all \( k \geq 0 \), we let \( z^{k+1} = Tz^k \).

Now, recall that \( z^* = 0 \). Thus, for all \( n \geq 0 \) and \( k \geq 0 \), we have

\[
\|z^k - z^*\|^2 = \|T^k z^0\|^2 = \sum_{i=0}^{\infty} b_i^{2(k+1)} \|z^0_i\|^2 = \sum_{i=0}^{\infty} \frac{b_i^{2(k+1)}}{(i + 1)^2} \geq \frac{b_n^{2(k+1)}}{(n + 1)^2}.
\]

Thus, \( \|z^k - z^*\| \geq b_n^{(k+1)}/(n + 1) \). To get the lower bound, we choose \( b_n \) and the sequence \((n_j)_{j \geq 0}\) using Corollary C.10 with any \( \eta \in (\max\{0, (a - 1)/(a + 1)\}, 1) \). Then we solve for the coefficients: \( c_n = 1 - \sqrt{(a + 1)(1 - b_n)/2} > 0 \).

**Remark C.4.** Theorems C.7 and C.9 show that the sequence \((z^j)_{j \geq 0}\) can converge arbitrarily slowly even if \((x^j_f)_{j \geq 0}\) and \((x^j_h)_{j \geq 0}\) converge with rate \( o(1/\sqrt{k+1}) \).
REFERENCES


[33] X. Cai, D. Han, and X. Yuan. The direct extension of ADMM for three-block separable convex minimization models is convergent when one function is strongly convex. Optimization Online, 2014. 229, 230


308


[83] M. Lichman. UCI machine learning repository, 2013. 251


