UNIVERSITY OF CALIFORNIA

Los Angeles

Operator Splitting Methods for Convex and Nonconvex Optimization

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

by

Yanli Liu

2020

ABSTRACT OF THE DISSERTATION

Operator Splitting Methods for Convex and Nonconvex Optimization

by

Yanli Liu
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 2020
Professor Wotao Yin, Chair

This dissertation focuses on a family of optimization methods called operator splitting methods. They solve complicated problems by decomposing the problem structure into simpler pieces and make progress on each of them separately. Over the past two decades, there has been a resurgence of interests in these methods as the demand for solving structured large-scale problems grew. One of the major challenges for splitting methods is their sensitivity to ill-conditioning, which often makes them struggle to achieve a high order of accuracy. Furthermore, their classical analyses are restricted to the nice settings where solutions do exist, and everything is convex. Much less is known when either of these assumptions breaks down.

This work aims to address the issues above. Specifically, we propose a novel acceleration technique called inexact preconditioning, which exploits second-order information at relatively low computation cost. We also show that certain splitting methods still work on problems without solutions, in the sense that their iterates provide information on what goes wrong and how to fix. Finally, for nonconvex problems with saddle points, we show that almost surely, splitting methods will only converge to the local minimums under certain assumptions.

The dissertation of Yanli Liu is approved.

Ali H. Sayed

Lieven Vandenberghe

Luminita Aura Vese

Wotao Yin, Committee Chair

University of California, Los Angeles

2020

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Wotao Yin, who has been a tremendous mentor during the past 5 years. Prof. Yin's high standards in both research and presentation have greatly improved the quality of my work, and have shaped me into an independent researcher. This dissertation would not have been possible without his guidance and encouragement.

I would also like to thank my the other members of my committee, Prof. Ali H. Sayed, Prof. Lieven Vandenberghe, and Prof. Luminita A. Vese for their advice and feedback. Their knowledge and insights have greatly enriched my work.

I am also indebted to my great collaborators: Prof. Ernest Ryu for his valuable suggestions on paper writing, and for his advice on doing research in general; Fei Feng, Robert Hannah, Yuejiao Sun and Yunbei Xu for their insights and hard work on our shared projects.

My friends and colleges have affected my Ph.D. career in a very positive way, including Qi Guo, Robert Hannah, Jialin Liu, Lecheng Ruan, Tao Sun, Baichuan Yuan, and Kun Yuan. I appreciate their advice on research and life in general.

Last but not the least, I am very grateful for my family back home for their unconditional love and support during this joyful journey.

Curriculum Vitae

2011 - 2015	Bachelor of Science in Physics, Minor Degree in Mathematics, Nankai University. Tianjin, China
2015 - 2020	Teaching and Research Assistant, Department of Mathematics, University of California, Los Angeles. Los Angeles, CA, USA
2018	Research Intern, Neo IVY Capital Management, New York, USA
2019	Research Intern, Alibaba DAMO Academy, Seattle, USA

Publications

Ryu, Ernest K., Yanli Liu, and Wotao Yin. "DouglasRachford splitting and ADMM for pathological convex optimization." Computational Optimization and Applications 74.3 (2019): 747-778.

Liu, Yanli, and Wotao Yin. "An envelope for DavisYin splitting and strict saddle-point avoidance." Journal of Optimization Theory and Applications 181.2 (2019): 567-587.

Liu, Yanli, Fei Feng, and Wotao Yin. "Acceleration of SVRG and Katyusha X by Inexact Preconditioning." International Conference on Machine Learning. 2019.

Liu, Yanli, Yunbei Xu, and Wotao Yin. "Acceleration of Primal-Dual Methods by Preconditioning and Simple Subproblem Procedures." arXiv preprint arXiv:1811.08937 (2018).

Hannah, Robert, Yanli Liu, Daniel O'Connor, and Wotao Yin. "Breaking the span assumption yields fast finite-sum minimization." Advances in Neural Information Pro-

cessing Systems. 2018.

Liu, Yanli, Ernest K. Ryu, and Wotao Yin. "A new use of DouglasRachford splitting for identifying infeasible, unbounded, and pathological conic programs." Mathematical Programming 177.1-2 (2019): 225-253.

Chang, Joshua C., Yanli Liu, and Tom Chou. "Reconstruction of Cell Focal Adhesions using Physical Constraints and Compressive Regularization." Biophysical journal 113.11 (2017): 2530-2539.

Part I

Introduction

CHAPTER 1

Introduction

This dissertation focuses on solving structured optimization problems. Specifically, we consider the composite optimization problem of the following form:

$$\text{minimize } f(x) + g(x), \tag{1.1}$$

where f and g are functions with special structures and can be nonsmooth or nonconvex. Common examples include the ℓ_1-norm $\|x\|_1$, the indicator function $\delta_C(x)$ of a nonempty convex set C, and the finite sum $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$. Formulation (1.1) captures problems in numerous research areas such as statistical and machine learning, medical imaging, compressed sensing, and control theory.

In these applications, the problem dimension (e.g., the size of the optimization variable x) is often very large, and the per-iteration computation cost of *second-order methods* (e.g., interior-point methods) is prohibitively high. Therefore, to solve (1.1) efficiently, *first-order methods* are widely applied. Among them, *operator splitting methods* are popular choices since they can decompose complicated problem structures into smaller pieces, and lead to simple algorithms that are easy to implement and have low per-iteration cost. This dissertation aims to accelerate the convergence of operator splitting methods and to analyze their behavior in the pathological[1] and nonconvex settings.

[1]Loosely speaking, pathological problems are the problems without a solution.

1.1 Background

Since the 1980s, a family of *second-order methods* called *interior-point methods* (IPMs) have been studied intensively by the optimization research community. In these methods, the key feature is the use of *logarithmic barrier function* to incorporate the problem constraints, and the logarithmic function serves well as a barrier function due to its *self-concordance property* [161]. IPMs were first applied for linear programs [98]. Soon after the important role of logarithmic barrier functions was understood, IPMs were generalized to quadratic and nonlinear problems [222]. Today, IPMs have been implemented efficiently in popular software packages [5, 166], and are often very suitable to solve problems of small or medium size. However, for many very large-scale problems that arise in modern machine learning, IPMs are often inefficient since at each iteration, IPMs require solving a linear system that involves second-order information of the objective. This linear system scales as the problem dimension and leads to a prohibitive per-iteration computation cost.

In view of this, much attention in optimization research has been directed to *first-order methods* in the past two decades. These methods only use first-order information such as gradient, subgradient, and proximal mapping[1], which are often cheap to obtain, and their cost scales well with the problem dimension. As a result, these methods are often easy to implement and enjoy low per-iteration cost. Two most prototypical examples are gradient descent and proximal point method [192], which only require evaluations of gradient or proximal mapping at each iteration. This makes them distinct from Newton's method, which is a classical second-order method. Aside from the aforementioned advantages, first-order methods are also amenable to parallelization, which is preferable for training large-scale models. One of the most notable examples is the widely applied Stochastic Gradient Descent (SGD) [189] for neural network training.

[1]The proximal mapping for a function f is defined as $\mathrm{Prox}_f(x) = \arg\min_y\{f(y) + \frac{1}{2}\|y - x\|^2\}$, a formal definition can be found at Sec. 1.4.

Based on gradient descent and proximal point method, numerous first-order algorithms have been developed. Some of the most popular prominent examples are, accelerated gradient methods [163, 29], stochastic gradient methods [189, 115], subgradient methods [205, 181, 205], mirror descent [159, 51, 28], coordinate descent [182, 113, 75], conditional gradient methods [93, 123, 79], and *operator splitting methods* that include, proximal gradient methods [133], Alternating Direction Method of Multipliers (ADMM) [94, 104], and Primal-Dual Hybrid Gradient (PDHG) [244, 49], and many other extensions. These methods are designed for different problem settings, but are related to each other in different ways. Furthermore, combining their features may lead to new algorithms that can solve more chanllenaging problems.

1.2 Operator Splitting Methods

Operator splitting methods are a family of first-order methods as they only rely on the first-order information of the objective. Their study originated from the seminal work by Sophus Lie on the Lie scheme in the 1890s [128]. At first, they were designed to solve the PDEs arising from computational physics. Later in the 1970s, the theory of monotone operators came into play, and these optimization methods were related to certain operator splitting schemes. For example, the projected gradient method corresponds to forward-backward splitting (FBS) [133], and the Alternating Direction Method of Multipliers (ADMM) corresponds to Douglas-Rachford splitting (DRS) [94]. This interpretation provides a unified view for these methods and has inspired the design and analysis of new optimization methods [69, 62, 235, 218, 45]. An overview of several common splitting methods can be found in Section 1.5.

The underlying principle of splitting methods is to decompose complicated problem structures into simple components, and deal with them separately by solving subproblems that only involve individual components. Another feature is that some of these components are allowed to be nonsmooth. In the past 15 years or so, optimization mod-

els in numerous research areas require solving nonsmooth optimization problems that are built up from simple components. To name a few, compressed sensing [76], Lasso [217], logistic regression [117] and image denoising [151] all involve sums of multiple functions and require an ℓ_1-norm penalty term to promote sparsity in their solutions. Therefore, the aforementioned advantages have led to the recent resurgence of interest in operator splitting methods.

However, splitting methods often suffer from their sensitivity to ill-conditioning, which is a common challenge for other first-order methods as well, due to the lack of second-order information. Furthermore, The classical analyses of splitting methods are constrained to non-pathological settings, where a primal-dual solution pair is assumed to exist, and strong duality holds. While in fact, even simple convex problems may not satisfy these assumptions[1]. Finally, the classical theory of splitting methods heavily relies on the monotonicity of the individual operators, which is lacking under nonconvex settings.

This dissertation aims to accelerate the convergence of operator splitting algorithms for convex problems and to analyze their behaviors in the pathological and nonconvex settings. The contributions are listed as follows.

1.3 Contributions

1.3.1 Acceleration by inexact preconditioning

As first-order algorithms, operator splitting methods suffer from slow (tail) convergence, especially on poorly conditioned problems. They may take thousands of iterations and still struggle to reach four digits of accuracy. While they have many advantages such as being easy to implement and friendly to parallelization, their sensitivity to problem conditions is their main disadvantage.

[1]For example, consider the following two problems: (i) $\min_{x\leq 1} x$, and (ii) Find $x \in [-2,-1]\cap[0,1]$.

To improve their performance of on ill-conditioned problems, researchers have tried to apply preconditioning, which is an idea first proposed for solving linear systems, and later applied to simple algorithms such as gradient descent. Recently, it has also been widely applied on splitting methods such as forward-backward splitting (FBS) [224, 61, 45, 230], Douglas-Rachford splitting (DRS) [42, 43], Primal-Dual Hybrid Gradient (PDHG) [179], and alternating directions method of multipliers (ADMM).[99, 100].

Depending on the application and how one applies splitting, these preconditioned algorithms may or may not have subproblems with closed-form solutions. When they do not, the cost of solving subproblems has to be taken into consideration. Previous works either assume the existence of an oracle that returns the exact solution of the subproblems, or allow approximate subproblem solutions with quickly diminishing errors [185, 83, 164, 126, 87, 86]. In either of these two cases, the total cost is prohibitive under realistic settings.

In Part II of this dissertation, we present a new preconditioning technique called *inexact preconditioning* and apply it to PDHG, ADMM, and Stochastic Variance-Reduced Gradient (SVRG). Conceptually, this technique involves two steps. First, one selects appropriate preconditioners based on specific problem structures and splitting algorithms. Then, one applies the preconditioned algorithms and solve the subproblems *highly inexactly* by *warmstart* and a *fixed* number of simple subroutines. Efficient subroutines can be chosen based on different subproblem structure and in particular, one does not need to enforce the errors to be diminishing in certain ways as in previous works. Theoretically, We show that this inexact preconditioning strategy brings significant acceleration to PDHG, ADMM, and SVRG. In practice, the efficacy of inexact preconditioning is demonstrated on several popular models such as logistic regression, graph cut, and computed tomography (CT) reconstruction, where a 4–95$\times$ speedup is observed.

1.3.2 Convergence behavior on pathological problems

Many convex optimization algorithms have strong theoretical guarantees and empirical performance, but they are often limited to non-pathological problems[1]; under pathologies often the theory breaks down and the empirical performance degrades significantly. In fact, the behavior of convex optimization algorithms under pathologies has been studied much less, and many existing solvers often simply report "failure" without informing the users of what went wrong upon encountering infeasibility, unboundedness, or other pathologies. Pathological problems are numerically challenging, but they are not impossible to deal with. As pathologies can arise in practice (see, for example, [141, 140, 225, 229, 78]), designing a robust algorithm that behaves well in all cases is important to the completion of a robust solver.

In Part III of this dissertation, we study the behavior of DRS and ADMM for pathological convex programs. Perhaps surprisingly, we show that although the iterates of DRS and ADMM diverge for pathological problems, the precise manner in which they diverge still provides useful information regarding the type of pathology that we encounter. Specifically, for a class of convex programs called *conic programming*, many pathologies can be identified by investigating the divergence pattern of the iterates. Furthermore, for certain types of pathologies, this divergence pattern informs us how to modify the pathological program to remove the pathology. For general convex problems, certain pathologies can still be identified, and we establish that DRS and ADMM only require strong duality to work even when the primal and/or dual solution does not exist, in the sense that the objective values of the iterates are asymptotically optimal.

[1] Problems that have both primal and dual solutions, and strongly duality holds.

1.3.3 Convergence behavior on nonconvex problems

Operator splitting methods are traditionally analyzed under the assumption that the subdifferentials of the objective functions are maximally monotone. While for nonconvex functions, their subdifferentials are generally non-monotone. Therefore, the majority of the existing results on splitting methods apply only to convex objective functions. Recently, FBS and DRS are found to numerically converge for certain nonconvex problems [209, 215, 125, 6, 54]. Theoretically, their iterates have been shown to converge to stationary points under some nonconvex settings [10, 125, 216, 108]. However, it remains possible that the limits of their convergent sequences are saddle points instead of local minimums.

In Part IV of this dissertation, we show that under some smoothness conditions, FBS and DRS can avoid the strict saddle points[1] almost surely, in the sense that the probability for DRS and FBS iterations with random initializations to converge to strict saddle points of their respective objectives is zero. The main technical tools to achieve this are (i) Forward-Backward Envelope (FBE) [215], Douglas-Rachford Envelope (DRE) [171] from nonconvex analysis, and (ii) Stable-Center Manifold Theorem [206] from dynamical systems.

FBE and DRE are functions with nice properties even in the nonconvex settings. In particular, they share the same stationary points, local minimizers, and strict saddle points with the objectives of FBS and DRS, respectively. Furthermore, the FBS and DRS iterations can be written as (preconditioned) gradient descent iterations on FBE and DRE. By analyzing these gradient descent iterations with the Stable-Center Manifold Theorem, one can show that whenever FBS and DRS converge, their limits will not be the strict saddles of FBE and DRE almost surely, which are exactly the strict saddles of their corresponding objective functions. Consequently, for many practical models that satisfy the *strict saddle property*[2], FBS and DRS will almost always avoid

[1] I.e., saddle points with a negative curvature.

the strict saddle points whenever they converge.

1.4 Notations and Preliminaries

In this section, we review standard notions of convex analysis, state several known results, and set up the notation. For the sake of brevity, we omit proofs or direct references of the standard results and refer interested readers to standard references such as [190, 195, 17]. Other relevant results will be provided at the beginning of each chapter.

We use $\|\cdot\|$ for ℓ_2-norm, $\|\cdot\|_1$ for ℓ_1-norm, and $\langle\cdot,\cdot\rangle$ for dot product. We use I_n to denote the identity matrix of size $n\times n$. $M\succ 0$ means M is a symmetric, positive definite matrix, and $M\succeq 0$ means M is a symmetric, positive semidefinite matrix.

We write $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ as the smallest and the largest eigenvalues of M, respectively, and $\kappa(M)=\frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$ as the condition number of M. For $M\succeq 0$, let $\|\cdot\|_M$ and $\langle\cdot,\cdot\rangle_M$ denote the semi-norm and inner product induced by M, respectively, i.e., $\langle x,y\rangle_M=x^TMy, \|x\|_M=\sqrt{x^TMx}$. If $M\succ 0$, then $\|\cdot\|_M$ is a norm.

A function f is convex if $f(\theta x+(1-\theta)y)\leq \theta f(x)+(1-\theta)f(y)$ for all $x,y\in\mathbb{R}^n$ and $\theta\in[0,1]$. A function f is closed if its epigraph $\left\{(x,\alpha)\in\mathbb{R}^{n+1}\mid f(x)\leq\alpha\right\}$ is a closed subset of $\mathbb{R}^{n+1}$. We say $f:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$ is proper if $f(x)<\infty$ for some x. In this work, we focus our attention on proper, closed, and convex (PCC) functions most of the time. If f and g are PCC functions, then $f+g$ is PCC or $f+g=\infty$ everywhere. If $\gamma>0$, then γf is PCC. Define the (effective) domain of f as $\mathbf{dom}\, f=\{x\in\mathbb{R}^n\,|\,f(x)<\infty\}$. For any $\gamma>0$, we have $\mathbf{dom}\,\gamma f=\mathbf{dom}\, f$.

For a proper closed convex function $f:\mathbb{R}^{\mathrm{n}}\to\mathbb{R}\cup\{+\infty\}$, its subdifferential at $x\in\mathbf{dom}\, f$ is written as

$$\partial f(x)=\{v\in\mathbb{R}^{\mathrm{n}}\mid f(z)\geq f(x)+\langle v,z-x\rangle,\ \forall z\in\mathbb{R}^{\mathrm{n}}\},$$

[2]That is, the stationary points of the objective are either local minimizers or strict saddle points.

and its convex conjugate as

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle y, x\rangle - f(x)\}.$$

We have $y \in \partial f(x)$ if and only if $x \in \partial f^*(y)$.

If f is convex and proper, then $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is PCC. If f is PCC, then $(f^*)^* = f$. For any $\gamma > 0$, we have $(\gamma f)^*(x) = \gamma f^*(x/\gamma)$ and $\mathbf{dom}\,(\gamma f)^* = \gamma \mathbf{dom}\, f^*$. If $h(x) = g(-x)$, then $h^*(y) = g^*(-y)$.

We say that $f : \mathbb{R}^{\mathrm{n}} \to \mathbb{R}$ is L_f-smooth, if it is differentiable and satisfies

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_f}{2}\|y - x\|^2, \forall x, y \in \mathbb{R}^{\mathrm{n}}\,.$$

Note that a smooth function f may be nonconvex.

We say that f is σ_f-strongly convex, if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\sigma_f}{2}\|y - x\|^2, \forall x, y \in \mathbb{R}^{\mathrm{n}}\,.$$

A set C is convex, if $x, y \in C$ and $\theta \in [0, 1]$ implies $\theta x + (1 - \theta)y \in C$. Write $\overline{C}$ for the closure of C. If C is convex $\overline{C}$ is convex. The Minkowski sum and differences of A and B are

$$A + B = \{a + b \,|\, a \in A,\, b \in B\}, \qquad A - B = \{a - b \,|\, a \in A,\, b \in B\},$$

respectively. If A and B are convex, then $A + B$ and $A - B$ are convex. However, neither $A + B$ nor $A - B$ is guaranteed to be closed, even when A and B are nonempty closed convex sets.

For the distance between $x \in \mathbb{R}^n$ and the set A, write

$$\mathrm{dist}(x, A) = \inf\{\|x - a\| \,|\, a \in A\}.$$

For the distance between A and B, write

$$\mathrm{dist}(A, B) = \inf\{\|a - b\| \,|\, a \in A,\, b \in B\}.$$

Note that $\text{dist}(A, B) = 0$ if and only if $\mathbf{0} \in \overline{A - B}$.

Define the projection onto C as $\Pi_C(x_0) = \arg\min_{x \in C} \|x - x_0\|$. When C is closed and convex, $\Pi_C : \mathbb{R}^n \to \mathbb{R}^n$ is well-defined, i.e., the minimizer uniquely exists.

Define the indicator function with respect to C as

$$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{otherwise.} \end{cases}$$

When C is closed convex, $\delta_C : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is PCC.

Define the support function of C as

$$\sigma_C(y) = \sup_{x \in C} \{\langle x, y \rangle\}.$$

$\sigma_C : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is PCC. When C is convex, we have $\sigma_C = \sigma_{\overline{C}}$. If A and B are convex, then $\sigma_{A+B} = \sigma_A + \sigma_B$. If C is closed and convex, then $(\sigma_C)^* = \delta_C$.

Define the proximal operator $\text{Prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ as

$$\text{Prox}_f(z) = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left\{ f(x) + (1/2)\|x - z\|^2 \right\}.$$

When f is PCC, the arg min uniquely exists, and therefore Prox_f is well-defined. When C is closed and convex, $\text{Prox}_{\delta_C} = \Pi_C$. When f is PCC, $\text{Prox}_f + \text{Prox}_{f^*} = I$, where $I : \mathbb{R}^n \to \mathbb{R}^n$ is the identity operator.

A mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive if $\|T(x) - T(y)\| \leq \|x - y\|$ for all $x, y \in \mathbb{R}^n$. Nonexpansive mappings are, by definition, Lipschitz continuous with Lipschitz constant 1. $T : \mathbb{R}^n \to \mathbb{R}^n$ is firmly-nonexpansive if

$$\|T(x) - T(y)\|^2 \leq \langle x - y, T(x) - T(y) \rangle$$

for all $x, y \in \mathbb{R}^n$. Proximal and projection operators are firmly-nonexpansive.

1.5 Common Operator Splitting Schemes

Now let us list some common operator splitting schemes. All of them can be cast as *fixed point iterations* of the form $z^{k+1} = Tz^k$, where T is a firmly-nonexpansive operator,

and z^k belongs to some Hilbert space $\mathcal{H}$.

Forward-backward splitting (FBS) [133] FBS solves the following problem:

$$\underset{x\in\mathcal{H}}{\text{minimize}}\, f(x)+g(x),$$

where f is PCC and smooth, and g is PCC.

Define $T=\text{Prox}_{\gamma g}(I-\gamma\nabla f)$. Then, the iteration of FBS can written as

$$x^{k+1}=Tx^k=\text{Prox}_{\gamma g}\left(x^k-\gamma\nabla f(x^k)\right).$$

Or equivalently,

$$\begin{aligned} y^k&=x^k-\gamma\nabla f(x^k),\\ x^{k+1}&=\text{Prox}_{\gamma g}(y^k), \end{aligned}$$

where $\gamma>0$ is a stepsize.

From the above iteration, we can see that FBS "splits" the problem by dealing with f and g separately.

Later, we will also work on another algorithm called **Stochastic Variance-Reduced Gradient (SVRG)** [114], which extends FBS to the following setting:

$$\underset{x\in\mathcal{H}}{\text{minimize}}\, f(x)+g(x)=\sum_{i=1}^{n}f_i(x)+g(x).$$

Here, f admits a finite sum structure, and its full gradient $\nabla f(x^k)$ may be expensive to obtain when n is large. In SVRG, a cheaper semi-stochastic gradient $\tilde{\nabla}^k$ at x^k is applied instead. Specifically,

$$\tilde{\nabla}^k=\nabla f(x^{k'})+\left(\nabla f_{i_k}(x^k)-\nabla f_{i_k}(x^{k'})\right),$$

where $\nabla f(x^{k'})$ is a previous full gradient at some iteration, and it will be recycled for some later iterations $k\geq k'$. i_k is picked uniformly at random from $\{1,2,...,n\}$.

SVRG iteration can be then written as

$$x^{k+1} = \mathrm{Prox}_{\gamma g}\left(x^k - \gamma\tilde{\nabla}^k\right).$$

Douglas-Rachford Splitting (DRS) [77] DRS solves the following problem

$$\underset{x\in\mathcal{H}}{\text{minimize}}\, f(x) + g(x),$$

where f and g are PCC.

Define $T = \frac{1}{2}I + \frac{1}{2}(2\,\mathrm{Prox}_{\gamma g} - I)(2\,\mathrm{Prox}_{\gamma f} - I)$, where $\gamma > 0$ is a stepsize. Then, DRS iteration can be written as

$$z^{k+1} = Tz^k,$$

or equivalently,

$$\begin{aligned} x^{k+1/2} &= \mathrm{Prox}_{\gamma f}(z^k),\\ x^{k+1} &= \mathrm{Prox}_{\gamma g}(2x^{k+1/2} - z^k),\\ z^{k+1} &= z^k + x^{k+1} - x^{k+1/2}. \end{aligned}$$

We will also work on another closely related algorithm called **Alternating Direction Method of Multipliers (ADMM)** [94, 104], which solves the following problem

$$\begin{aligned} \underset{x\in\mathbb{R}^p, y\in\mathbb{R}^q}{\text{minimize}} \quad & f(x) + g(y)\\ \text{subject to} \quad & Ax + By = c, \end{aligned}$$

where $f : \mathbb{R}^p \to \mathbb{R}\cup\{\infty\}$ and $g : \mathbb{R}^q \to \mathbb{R}\cup\{\infty\}$ are PCC, $A \in \mathbb{R}^{n\times p}$, $B \in \mathbb{R}^{n\times q}$, and $c \in \mathbb{R}^n$,

$$\begin{aligned} x^{k+1} &\in \underset{x\in\mathbb{R}^p}{\arg\min}\left\{f(x) + \langle\nu^k, Ax + By^k - c\rangle + \frac{1}{2\gamma}\|Ax + By^k - c\|^2\right\}\\ y^{k+1} &\in \underset{y\in\mathbb{R}^q}{\arg\min}\left\{g(y) + \langle\nu^k, Ax^{k+1} + By - c\rangle + \frac{1}{2\gamma}\|Ax^{k+1} + By - c\|^2\right\}\\ \nu^{k+1} &= \nu^k + (1/\gamma)(Ax^{k+1} + By^{k+1} - c). \end{aligned}$$

Under mild regularity assumptions, it can be shown that DRS and ADMM are equivalent [82, 84, 236].

Primal-Dual Hybrid Gradient (PDHG) [49] The method of Primal-Dual Hybrid Gradient or PDHG for solving solves

$$\underset{x\in\mathbb{R}^{\mathrm{n}}}{\text{minimize}}\, f(x)+g(Ax),$$

where $f:\mathbb{R}^{\mathrm{n}}\to\mathbb{R}\cup\{\infty\}$, $g:\mathbb{R}^{\mathrm{m}}\to\mathbb{R}\cup\{\infty\}$ are PCC, and $A\in\mathbb{R}^{m\times n}$. PDHG refers to the iteration

$$\begin{aligned} x^{k+1} &= \mathrm{Prox}_{\tau f}(x^k-\tau A^T z^k),\\ z^{k+1} &= \mathrm{Prox}_{\sigma g^*}(z^k+\sigma A(2x^{k+1}-x^k)), \end{aligned}$$

where $\tau,\sigma>0$ are stepsizes.

Define

$$A=\begin{pmatrix}\partial f & A^T\\ -A & \partial g^*\end{pmatrix},$$

and let

$$M=\begin{pmatrix}\frac{1}{\tau}I_n & -A^T\\ -A & \frac{1}{\sigma}I_m\end{pmatrix}\succ 0.$$

Then, the above PDHG iteration can be written as

$$y^{k+1}=Ty^k=(I+M^{-1}A)^{-1}y^k,$$

where $y^k=(x^k,z^k)^T$.

The operator $T=(I+M^{-1}A)^{-1}$ is firmly nonexpansive in $\|\cdot\|_M$, and y^k will converge to a primal-dual solution pair [112].

Part II

Acceleration by Inexact Preconditioning

In this part, we present the *inexact preconditioning* technique for accelerating several operator splitting algorithms, the results can also be found in [138] and [136].

The inexact preconditioning technique consists of two steps: (i) find appropriate preconditioner(s) based on the objective and the specific splitting algorithm, and (ii) solve the subproblems inexactly by just a *fixed* number of simple subroutines.

In Chapter 2, we apply this technique to accelerate PDHG and ADMM, the resulting algorithm is called inexact preconditioned PDHG (iPrePDHG), which is summarized in Algorithm 2.1. First, we provide a criterion for choosing preconditioners in Lemma 2.3.1 and Theorem 2.3.2. It turns out that most of the time, the optimal preconditioners will be non-diagonal, and the subproblems will not have closed-form solutions. Therefore, we propose to solve them until a certain condition is satisfied (see Definition 2.3.1). Remarkably, this condition is easily satisfied by applying some simple subroutines a fixed number of times (see Theorems 2.3.3 and 2.3.4). Finally, we prove the global convergence of iPrePDHG in Theorem 2.3.9, and provide extensive numerical tests in Section 2.4.

The structure of Chapter 3 is similar. We aim to accelerate SVRG and Katyusha X[1] by inexact preconditioning, and the new algorithms are called iPreSVRG and iPreKatX, respectively (see Algorithms 3.1 and 3.2). The preconditioner M should decrease the condition number and can vary for different objectives(see Definition 3.2.3). in Section 3.5. To prove acceleration, we first show that it is fine to solve the subproblems by applying FISTA with restart a small number of times so that a certain error condition will be satisfied (see Lemmas 3.4.1 and 3.4.4). Furthermore, when this error condition is satisfied, the global convergence of iPreSVRG and iPreKatX is guaranteed (see Theorems 3.4.2 and 3.4.3). Finally, we proved the acceleration of iPreSVRG over SVRG, as well as the acceleration of iPreKatX over Katyusha X in Theorems 3.4.5 and 3.4.6. This acceleration is also observed numerically on Lasso and logistic regression.

[1]Katyusha X is a Nesterov-accelerated version of SVRG.

CHAPTER 2

Inexact Preconditioning for PDHG and ADMM

2.1 Introduction

In this chapter, we consider the following optimization problem:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\, f(x) + g(Ax), \tag{2.1}$$

together with its dual problem:

$$\underset{z\in\mathbb{R}^m}{\text{minimize}}\, f^*(-A^T z) + g^*(z), \tag{2.2}$$

where $f : \mathbb{R}^{\mathrm{n}} \to \mathbb{R}\cup\{+\infty\}$ and $g : \mathbb{R}^{\mathrm{m}} \to \mathbb{R}\cup\{+\infty\}$ are closed proper convex, and $A \in \mathbb{R}^{m\times n}$ is a matrix, f^* and g^* are the convex conjugates of f and g, respectively.

Formulations (2.1) or (2.2) are abstractions of many application problems, which include image restoration [244], magnetic resonance imaging [221], network optimization [90], computer vision [180], and earth mover's distance [127]. For many of them, primal-dual algorithms such as Primal-Dual Hybrid Gradient (PDHG) and Alternating Direction Method of Multipliers (ADMM) have been popular choices.

However, as a first-order algorithm, PDHG and ADMM suffer from slow (tail) convergence especially on poorly conditioned problems. They may take thousands of iterations and still struggle reaching just four digits of accuracy. While they have many advantages such as being easy to implement and friendly to parallelization, their sensitivity to problem conditions is their main disadvantage.

To improve the performance of PDHG and ADMM, researchers have tried using preconditioners, which has been widely applied for forward-backward type of methods

[224, 61, 45], as well as other methods [44, 62, 112, 223]. Depending on the application and how one applies splitting, preconditioned PDHG and ADMM may or may not have subproblems with closed-form solutions. When they do not, researchers have studied approximate subproblem solutions to reduce the total running time. In this work, we propose a new way of applying preconditioning that outperforms the existing state-of-the-art.

2.1.1 Proposed approach

Simply speaking, we find a way to have both non-diagonal preconditioners (thus much fewer iterations) and very simple subproblem procedures (thus maintaining the advantages of PDHG and ADMM).

First, we apply preconditioning. We present Preconditioned PDHG (PrePDHG) along with its convergence condition and a performance bound. We propose to choose preconditioners to optimize the bound. In the special case where one preconditioner is trivially fixed as an identity matrix, optimizing the bound gives us the optimal choice of the other preconditioner, which actually reduces PrePDHG to ADMM. This observation explains why ADMM often takes fewer iterations than PDHG (as PDHG sets both preconditioners to identity matrices).

Next, we study how to solve PrePDHG subproblems. In all applications we are aware of, only one of the two subproblem is (subject to) ill-conditioned. (After all, we can always apply splitting to gether ill-conditioned components into one subproblem.) Therefore, we choose a non-diagonal preconditioner for the ill-conditioned subproblem and a trivial or diagonal preconditioner for the other subproblem. Again, the pair of preconditioners should be chosen to (nearly) optimize the performance bound. Since the non-diagonal preconditioner introduces dependence between different coordinates, its subproblem generally does not have a closed-form solution. In particular, if the subproblem has an ℓ_1-norm, which is often the reason why PDHG or ADMM is used, it

often loses its closed-form solution due to the preconditioner. Therefore, we propose to approximately solve it to satisfy an accuracy condition. Remarkably, there is no need to dynamically stops a subproblem procedure to honor the condition. Instead, the condition is automatically satisfied as long as one applies a common iterative procedure for some *fixed number* of iterations, which is new in the literature. Common choices of the procedure include proximal gradient descent, FISTA with restart, proximal block coordinate descent, and accelerated block-coordinate-gradient-descent (BCGD) methods (e.g., [132, 3, 109]). We call this method iPrePDHG (i for "inexact").

Next, we establish the overall convergence of iPrePDHG. To handle the inexact subproblem, we first transform iPrePDHG into an equivalent form and then analyze an Lyapunov function to establish convergence. The technique in our proof appears to be new in the PDHG and ADMM literature.

Finally, we apply our approach on a few applications including image denoising, graph cut, optimal transport, and CT reconstruction. For the last application, we use a diagonal preconditioner in one subproblem, which gives it a closed-form solution, and a non-diagonal preconditioner in the other, which we approximately solve. In each of the other applications, one subproblem uses no (identity) preconditioner, and the other uses a non-diagonal preconditioner. We numerically evaluated the performance of iPrePDHG using these recommended preconditioners and observed speedups of 4–95 times over the existing state-of-the-art.

Since we show ADMM is a special PrePDHG with one trivial preconditioner, our approach can also accelerate ADMM. In fact, for three of the above four applications, there are one trivial preconditioner in each, so their iPrePDHG are inexact preconditioned ADMM.

2.1.2 Related Literature

Many problems to which we apply PDHG have separable functions f or g, or both, so the resulting PDHG subproblems often (though not always) have closed-form solutions. When subproblems are simple, we care mainly about the convergence rate of PDHG, which depends on the problem conditioning. To accelerate PDHG, diagonal preconditioning [179] was proposed since its diagonal structure maintains closed-form solutions for the subproblems and, therefore, reduces iteration complexity without making each iteration more difficult. In comparison, non-diagonal preconditioners are much more effective at reducing iteration complexity, but their off-diagonal entries couple different components in the subproblems, causing the lost of closed-form solutions of subproblems.

When a PDHG subproblem has no closed-form solution, one often uses an iterative algorithm to approximately solve it. We call it Inexact PDHG. Under certain conditions, Inexact PDHG still converges to the exact solution. Specifically, [185] uses three different types of conditions to skillfully control the errors of the subproblems; all those errors need to be summable over all the iterations and thereby requiring the error to diminish asymptotically. In an interesting method from [42, 43], one subproblem computes a proximal operator of a convex quadratic function, which can include a preconditioner and still has a closed-form solution involving matrix inversion. This proximal operator is successively applied n times in each iteration, for $n \geq 1$.

ADMM has different subproblems. One of its subproblems minimizes the sum of $f(x)$ and a squared term involving Ax. Only when A has special structures does the subproblem have closed-form solutions. Inexact ADMM refers to the ADMM with at least one of its subproblems inexactly solved. An *absolute error criterion* was introduced in [83], where the subproblem errors are controlled by a summable (thus diminishing) sequence of error tolerances. To simplify the choice of the sequences, a *relative error criterion* was adopted in several later works, where the subproblem errors are controlled

by a single parameter multiplying certain quantities that one can compute during the iterations. In [164], the parameters need to be square summable. In [126], the parameters are constants when both objective functions are Lipschitz differentiable. In [87, 86], two possible outcomes of the algorithm are described: (i) infinite outer loops and finite inner loops, and (ii) finite outer loops and the last inner loop is infinite, both guaranteeing convergence to a solution. On the other hand, it is unclear how to recognize them. Since there is no bound on the number of inner loops in case (i), one may recognize it as case (ii) and stop the algorithm before it converges.

There are works that apply certain kinds of preconditioning to accelerate ADMM. Paper [99] uses diagonal preconditioning and observes improved performance. After that, non-diagonal preconditioning is analyzed [42, 43], which presents effective preconditioners for specific applications. One of their preconditioners needs to be inverted (though not needed in our method). Recently, preconditioning for problems with linear convergence has also been studied with promising numerical performances [100].

2.1.3 Organization

The rest of this chapter is organized as follows: Section 2.2 establishes notation and reviews basics. In the first part of Section 2.3, we provide a criterion for choosing preconditioners. In its second part, we introduce the condition for inexact subproblems, which can be automatically satisfied by iterating a fixed number of certain inner loops. This method is called iPrePDHG. In the last part of Section 2.3, we establish the convergence of iPrePDHG. Section 2.4 describes specific preconditioners and reports numerical results. Finally, Section 2.5 concludes this chapter.

2.2 Preliminaries

In addition to the preliminaries introduced in Sec. 1.4, we need the following in this chapter.

For any $M \succ 0$, we define the extended proximal operator of ϕ as

$$\mathrm{Prox}_{\phi}^{M}(x) := \arg\min_{y \in \mathbb{R}^n} \{\phi(y) + \frac{1}{2}\|y - x\|_M^2\}. \tag{2.3}$$

If $M = \gamma^{-1} I$ for $\gamma > 0$, it reduces to a classic proximal operator.

We also have the following generalization of Moreau's Identity:

Lemma 2.2.1 ([60], Theorem 3.1(ii))**.** *For any proper closed convex function ϕ and $M \succ 0$, we have*

$$x = \mathrm{Prox}_{\phi}^{M}(x) + M^{-1}\,\mathrm{Prox}_{\phi^*}^{M^{-1}}(Mx). \tag{2.4}$$

We say a proper closed function ϕ is a Kurdyka-ojasiewicz (KL) function if, for each $x_0 \in \mathbf{dom}\phi$, there exist $\eta \in (0, \infty]$, a neighborhood U of x_0, and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ such that:

1. $\varphi(0) = 0$,
2. φ is C^1 on $(0, \eta)$,
3. for all $s \in (0, \eta), \varphi'(s) > 0$,
4. for all $x \in U \cap \{x \,|\, \phi(x_0) < \phi(x) < \phi(x_0) + \eta\}$, the KL inequality holds:

$$\varphi'(\phi(x) - \phi(x_0))\mathrm{dist}(0, \partial\phi(x)) \geq 1.$$

2.3 Main results

This section presents the key results of this chapter. In Sec. 2.3.1 we demonstrate how to apply preconditioning to PDHG. Then, we establish rules of preconditioner selection

in Sec. 2.3.2. In Sec. 2.3.3, we present the proposed method iPrePDHG. Finally, we establish the convergence of iPrePDHG in Sec. 2.3.4.

Throughout this section, we assume the following regularity assumptions:

Assumption 2.3.1.

1. $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ *are proper closed convex.*
2. *A primal-dual solution pair* $(x^\star, z^\star)$ *of* (2.1) *and* (2.2) *exists, i.e.,*

$$\mathbf{0} \in \partial f(x^\star) + A^T z^\star, \quad \mathbf{0} \in \partial g(Ax^\star) - z^\star.$$

The problem (2.1) also has the following convex-concave saddle-point formulation:

$$\min_{x\in\mathbb{R}^n} \max_{z\in\mathbb{R}^m} \varphi(x, z) \coloneqq f(x) + \langle Ax, z\rangle - g^*(z). \tag{2.5}$$

A primal-dual solution pair $(x^\star, z^\star)$ is a solution of (2.5).

2.3.1 Preconditioned PDHG

The method of Primal-Dual Hybrid Gradient or PDHG [244, 49] for solving (2.1) refers to the iteration

$$\begin{aligned} x^{k+1} &= \mathrm{Prox}_{\tau f}(x^k - \tau A^T z^k), \\ z^{k+1} &= \mathrm{Prox}_{\sigma g^*}(z^k + \sigma A(2x^{k+1} - x^k)). \end{aligned} \tag{2.6}$$

When $\frac{1}{\tau\sigma} \geq \|A\|^2$, the iterates of (2.6) converge [49] to a primal-dual solution pair of (2.1). We can generalize (2.6) by applying preconditioners $M_1, M_2 \succ 0$ (their choices are discussed below) to obtain Preconditioned PDHG or PrePDHG:

$$\begin{aligned} x^{k+1} &= \mathrm{Prox}_f^{M_1}(x^k - M_1^{-1} A^T z^k), \\ z^{k+1} &= \mathrm{Prox}_{g^*}^{M_2}(z^k + M_2^{-1} A(2x^{k+1} - x^k)), \end{aligned} \tag{2.7}$$

where the extended proximal operators $\mathrm{Prox}_f^{M_1}$ and $\mathrm{Prox}_{g^*}^{M_2}$ are defined in (2.3). We can obtain the convergence of PrePDHG using the analysis in [50].

There is no need to compute M_1^{-1} and M_2^{-1} since (2.7) is equivalent to

$$\begin{aligned} x^{k+1} &= \arg\min_{x\in\mathbb{R}^n}\{f(x) + \langle x - x^k, A^T z^k\rangle + \frac{1}{2}\|x - x^k\|_{M_1}^2\}, \\ z^{k+1} &= \arg\min_{z\in\mathbb{R}^m}\{g^*(z) - \langle z - z^k, A(2x^{k+1} - x^k)\rangle + \frac{1}{2}\|z - z^k\|_{M_2}^2\}. \end{aligned} \tag{2.8}$$

2.3.2 Choice of preconditioners

In this section, we discuss how to select appropriate preconditioners M_1 and M_2. As a by-product, we show that ADMM corresponds to choosing $M_1 = \frac{1}{\tau} I_n$ and optimally choosing $M_2 = \tau AA^T$, thereby, explaining why ADMM appears to be faster than PDHG.

Let us start with the following lemma, which characterizes primal-dual solution pairs of (2.1) and (2.2).

Lemma 2.3.1. *Under Assumption 2.3.1, (X, Z) is a primal-dual solution pair of (2.1) if and only if $\varphi(X, z) - \varphi(x, Z) \le 0$ for any $(x, z) \in \mathbb{R}^{n+m}$, where φ is given in the saddle-point formulation (2.5).*

Proof. If (X, Z) is a primal-dual solution pair of (2.1), then

$$-A^T Z \in \partial f(X), \quad AX \in \partial g^*(Z).$$

Hence, for any $(x, z) \in \mathbb{R}^{n+m}$ we have

$$f(x) \ge f(X) + \langle -A^T Z, x - X\rangle, \quad g^*(z) \ge g^*(Z) + \langle AX, z - Z\rangle.$$

Adding them together yields $\varphi(X, z) - \varphi(x, Z) \le 0$.

On the other hand, if $\varphi(X, z) - \varphi(x, Z) \le 0$ for any $(x, z) \in \mathbb{R}^{n+m}$, then

$$\langle AX, z\rangle + f(X) - g^*(z) - \langle Ax, Z\rangle - f(x) + g^*(Z) \le 0 \quad \text{for any } (x, z) \in \mathbb{R}^{n+m}.$$

Taking $x = X$ yields $\langle AX, z - Z\rangle - g^*(z) + g^*(Z) \le 0$, so $AX \in \partial g^*(Z)$; Similarly, taking $z = Z$ gives $\langle AX - Ax, Z\rangle + f(X) - f(x) \le 0$, so $-A^T Z \in \partial f(X)$. As a result, (X, Z) is a primal-dual solution pair of (2.1). □

We present the following convergence result, adapted from Theorem 1 of [50].

Theorem 2.3.2. *Let $(x^k, z^k), k = 0, 1, ..., N$ be a sequence generated by PrePDHG (2.7). Under Assumption 2.3.1, if in addition*

$$\tilde{M} := \begin{pmatrix} M_1 & -A^T \\ -A & M_2 \end{pmatrix} \succeq 0, \tag{2.9}$$

then, for any $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$, it holds that

$$\varphi(X^N, z) - \varphi(x, Z^N) \leq \frac{1}{2N}(x - x^0, z - z^0) \begin{pmatrix} M_1 & -A^T \\ -A & M_2 \end{pmatrix} \begin{pmatrix} x - x^0 \\ z - z^0 \end{pmatrix}, \tag{2.10}$$

where $X^N = \frac{1}{N}\sum_{i=1}^N x^i$ and $Z^N = \frac{1}{N}\sum_{i=1}^N z^i$.

Proof. This follows from Theorem 1 of [50] by setting $L_f = 0$, $\frac{1}{\tau}D_x(x, x_0) = \frac{1}{2}\|x - x^0\|^2_{M_1}$, $\frac{1}{\sigma}D_z(z, z_0) = \frac{1}{2}\|z - z^0\|^2_{M_2}$, and $K = A$. Note that in Remark 1 of [50], D_x and D_z need to be 1−strongly convex to ensure their inequality (13) holds, which is exactly our (2.9). Therefore, we do not need D_x and D_z to be strongly convex. □

Based on the above results, one approach to accelerate convergence is to choose preconditioners M_1 and M_2 to obey (2.9) and minimize the right-hand side of (2.10). When a pair of preconditioner matrices attains this minimum, we say they are optimal. When one of them is fixed, the other that attains the minimum is also called optimal.

By Schur complement, the condition (2.9) is equivalent to $M_2 \succeq AM_1^{-1}A^T$. Hence, for any given $M_1 \succ 0$, the optimal M_2 is $AM_1^{-1}A^T$.

Original PDHG (2.6) corresponds to $M_1 = \frac{1}{\tau}I_n$, $M_2 = \frac{1}{\sigma}I_m$ with τ and σ obeying $\frac{1}{\tau\sigma} \geq \|A\|^2$ for convergence. In Appendix 2.A, we show that ADMM for problem (2.1) corresponds to setting $M_1 = \frac{1}{\tau}I_n$, $M_2 = \tau AA^T$, M_2 is optimal since $AM_1^{-1}A^T = \tau AA^T = M_2$ (This is related to, but different from, the result in [49, Sec. 4.3] stating that PDHG is equivalent to a preconditioned ADMM). In the next section, we show that when the z−subproblem is solved inexactly, a choice of $M_1 = \frac{1}{\tau}I_n, M_2 = \tau AA^T + \theta I_m$ with a small θ guarantees convergence (see Proposition 2.3.7).

By using more general pairs of M_1, M_2, we can potentially have even fewer iterations of PrePDHG than ADMM.

2.3.3 PrePDHG with fixed inner iterations

It wastes total time to solve the subproblems in (2.8) very accurately. It is more efficient to develop a proper condition and stop the subproblem procedure, which we call *inner iterations*, once the condition is satisfied. It is even better if we can simply fix the number of inner iterations and still guarantee global convergence.

In this subsection, we describe the "bounded relative error" of the z-subproblem in (2.7) and then show that this can be satisfied by running a fixed number of inner iterations, uniformly for every outer loop, which is new in the literature.

Definition 2.3.1 (Bounded relative error condition)**.** *Given x^k, x^{k+1} and z^k, we say that the z-subproblem in PrePDHG (2.7) is solved to a bounded relative error by some iterator S, if there is a constant $c > 0$ such that*

$$\mathbf{0} \in \partial g^*(z^{k+1}) + M_2\Big(z^{k+1} - z^k - M_2^{-1}A(2x^{k+1} - x^k)\Big) + \varepsilon^{k+1}, \tag{2.11}$$

$$\|\varepsilon^{k+1}\| \le c\|z^{k+1} - z^k\|. \tag{2.12}$$

Remarkably, this condition does not need to be checked at run time. For a fixed $c > 0$, the condition can be satisfied by a fixed number of inner iterations using, for example, S being the proximal gradient iteration (Theorem 2.3.3). One can also use faster solvers, e.g., FISTA with restart [167], and solvers that suit the subproblem structure, e.g., cyclic proximal BCD (Theorem 2.3.4). Although the error in solving z-subproblems appears to be neither summable nor square summable, convergence can still be established. But first, we summarize this method in Algorithm 2.1.

Theorem 2.3.3. *Take Assumption 2.3.1. Suppose in iPrePDHG, or Algorithm 2.1, we choose S as the proximal-gradient step with stepsize $\gamma \in (0, \frac{2\lambda_{\min}(M_2)}{\lambda^2_{\max}(M_2)})$ and repeat it p*

Algorithm 2.1 Inexact Preconditioned PDHG or iPrePDHG

Input: f, g, A in (2.1), preconditioners M_1 and M_2, initial (x_0, z_0), z-subproblem iterator S, inner iteration number p, max outer iteration number K.

Output: (x^K, z^K)

1: **for** $k \leftarrow 0, 1, ..., K-1$ **do**
2: $\quad x^{k+1} = \text{Prox}_f^{M_1}(x^k - M_1^{-1}A^T z^k)$;
3: $\quad z_0^{k+1} = z^k$;
4: $\quad$ **for** $i \leftarrow 0, 1, ..., p-1$ **do**
5: $\quad\quad z_{i+1}^{k+1} = S(z_i^{k+1}, x^{k+1}, x^k)$;
6: $\quad$ **end for**
7: $\quad z^{k+1} = z_p^{k+1}$; $\quad\quad \triangleright$ which approximates $\text{Prox}_{g^*}^{M_2}(z^k + M_2^{-1}A(2x^{k+1} - x^k))$
8: **end for**

times, where $p \geq 1$. Then, $z^{k+1} = z_p^{k+1}$ is an approximate solution to the z-subproblem up to a bounded relative error in Definition 2.3.1 for

$$c = c(p) = \frac{\frac{1}{\gamma} + \lambda_{\max}(M_2)}{1 - \rho^p}(\rho^p + \rho^{p-1}), \tag{2.13}$$

where $\rho = \sqrt{1 - \gamma(2\lambda_{\min}(M_2) - \gamma\lambda_{\max}^2(M_2))} < 1$.

Proof. The z-subproblem in (2.8) is of the form

$$\underset{z \in \mathbb{R}^m}{\text{minimize}}\, h_1(z) + h_2(z), \tag{2.14}$$

for $h_1(z) = g^*(z)$ and $h_2(z) = \frac{1}{2}\|z - z^k - M_2^{-1}A(2x^{k+1} - x^k)\|_{M_2}^2$. With our choice of S as the proximal-gradient descent step, the inner iterations are

$$\begin{aligned} z_0^{k+1} &= z^k, \\ z_{i+1}^{k+1} &= \text{Prox}_{\gamma h_1}(z_i^{k+1} - \gamma\nabla h_2(z_i^{k+1})), \quad i = 0, 1, ..., p-1, \end{aligned} \tag{2.15}$$

Concerning the last iterate $z^{k+1} = z_p^{k+1}$, we have from the definition of $\text{Prox}_{\gamma h_1}$ that

$$\mathbf{0} \in \partial h_1(z_p^{k+1}) + \nabla h_2(z_{p-1}^{k+1}) + \frac{1}{\gamma}(z_p^{k+1} - z_{p-1}^{k+1}).$$

Compare this with (2.11) and use $z^{k+1} = z_p^{k+1}$ to get

$$\varepsilon^{k+1} = \frac{1}{\gamma}(z_p^{k+1} - z_{p-1}^{k+1}) + \nabla h_2(z_{p-1}^{k+1}) - \nabla h_2(z_p^{k+1}).$$

It remains to show that ε^{k+1} satisfies (2.12).

Let $z_\star^{k+1}$ be the solution of (2.14), $\alpha = \lambda_{\min}(M_2)$, and $\beta = \lambda_{\max}(M_2)$. Then $h_1(z)$ is convex and $h_2(z)$ is α-strongly convex and β-Lipschitz differentiable. Consequently, [18, Prop. 26.16(ii)] gives

$$\|z_i^{k+1} - z_\star^{k+1}\| \leq \rho^i \|z_0^{k+1} - z_\star^{k+1}\|, \quad \forall i = 0, 1, ..., p,$$

where $\rho = \sqrt{1 - \gamma(2\alpha - \gamma\beta^2)}$.

Let $a_i = \|z_i^{k+1} - z_\star^{k+1}\|$. Then, $a_i \leq \rho^i a_0$. We can derive

$$\|\varepsilon^{k+1}\| \leq (\frac{1}{\gamma} + \beta)\|z_p^{k+1} - z_{p-1}^{k+1}\| \leq (\frac{1}{\gamma} + \beta)(a_p + a_{p-1}) \leq (\frac{1}{\gamma} + \beta)(\rho^p + \rho^{p-1})a_0. \quad (2.16)$$

On the other hand, we have

$$\|z^{k+1} - z^k\| \geq a_0 - a_p \geq (1 - \rho^p)a_0. \quad (2.17)$$

Combining these two equations yields

$$\|\varepsilon^{k+1}\| \leq c\|z^{k+1} - z^k\|,$$

where c is given in (2.13). □

Theorem 2.3.3 uses the iterator S that is the proximal-gradient step. It is straightforward to extend its proof to S being the FISTA step with restart. We omit the proof.

In our next theorem, we let S be the iterator of one epoch of the cyclic proximal BCD method. A BCD method updates one block of coordinates at a time while fixing the remaining blocks. In one epoch of cyclic BCD, all the blocks of coordinates are sequentially updated, and every block is updated once. In cyclic *proximal* BCD, each

block of coordinates is updated by a proximal-gradient step, just like (2.15) except only the chosen block is updated each time. When h_1 is block separable, each update costs only a fraction of updating all the blocks together. When different blocks are updated one after another, the Gauss-Seidel effect brings more progress. In addition, since the Lipschitz constant of each block gradient of h_2 is typically less than than that of ∇h_2, one can use a larger stepsize γ and get potentially even faster progress. Therefore, the iterator of cyclic proximal BCD is a better choice for S.

In summary, with $h_1(z) = g^*(z)$ and $h_2(z) = \frac{1}{2}\|z - z^k - M_2^{-1}A(2x^{k+1} - x^k)\|_{M_2}^2$, one epoch of cyclic proximal BCD for the $z-$subproblem can be written as

$$\begin{aligned} z_0^{k+1} &= z^k, \\ z_{i+1}^{k+1} &= S(z_i^{k+1}, x^{k+1}, x^k), \quad i = 0, 1, ..., p-1, \\ z^{k+1} &= z_p^{k+1}. \end{aligned}$$

where S is the iterator of cyclic proximal BCD. Define

$$\begin{aligned} T(z) &= \mathrm{Prox}_{\gamma h_1(z)}(z - \gamma \nabla h_2(z)), \\ B(z) &= \frac{1}{\gamma}(z - T(z)), \end{aligned}$$

and the jth coordinate operator of B:

$$B_j(z) = (0, ..., (B(z))_j, ..., 0), \quad j = 1, 2, ..., l.$$

Then, we have

$$z_{i+1}^{k+1} = S(z_i^{k+1}, x^{k+1}, x^k) = (I - \gamma B_l)(I - \gamma B_2)...(I - \gamma B_1) z_i^{k+1}.$$

Theorem 2.3.4. *Let Assumption 2.3.1 hold and g be block separable, i.e., $z = (z_1, z_2, ..., z_l)$ and $g(z) = \sum_{j=1}^{l} g_j(z_j)$. Suppose in iPrePDHG, or Algorithm 2.1, we choose S as the iterator of cyclic proximal BCD with stepsize γ satisfying*

$$0 < \gamma \leq \min \left\{ \frac{2\lambda_{\min}(M_2))}{\lambda_{\max}^2(M_2))}, \frac{1 - \sqrt{1 - \gamma(2\lambda_{\min}(M_2) - \gamma\lambda_{\max}^2(M_2))}}{4\sqrt{2}\gamma l \lambda_{\max}(M_2)}, \right.$$
$$\left. \frac{1}{4l\lambda_{\max}(M_2)}, \frac{2l\lambda_{\max}(M_2)}{17l\lambda_{\max}(M_2) + 2(\frac{1-\sqrt{1-\gamma(2\lambda_{\min}(M_2)-\gamma\lambda_{\max}^2(M_2))}}{\gamma})^2} \right\},$$

and we set $p \geq 1$. *Then,* $z^{k+1} = z_p^{k+1}$ *is an approximate solution to the* z*-subproblem up to a bounded relative error in Definition 2.3.1 for*

$$c = c(p) = \frac{(l\lambda_{\max}(M_2) + \frac{1}{\gamma})(\rho^p + \rho^{p-1})}{1 - \rho^p}, \tag{2.18}$$

where $\rho = 1 - \frac{\left(1-\sqrt{1-\gamma(2\lambda_{\min}(M_2)-\gamma\lambda_{\max}^2(M_2))}\right)^2}{2\gamma} < 1.$

Proof. See Appendix 2.B. □

2.3.4 Global convergence of iPrePDHG

In this subsection, we proceed to establishing the convergence of Algorithm 2.1. Our approach first transforms Algorithm 2.1 into an equivalent algorithm in Proposition 2.3.5 below and then proves its convergence in Theorems 2.3.8 and 2.3.9 below.

First, let us show that PrePDHG (2.7) is equivalent to an algorithm applied on the dual problem (2.2). This equivalence is analogous to the equivalence between PDHG (2.6) and Linearized ADMM applied to the dual problem (2.2), shown in [88]). Specifically, PrePDHG is equivalent to

$$\begin{aligned} z^{k+1} &= \mathrm{Prox}_{g^*}^{M_2}(z^k + M_2^{-1}AM_1^{-1}(-A^Tz^k - y^k + u^k)), \\ y^{k+1} &= \mathrm{Prox}_{f^*}^{M_1^{-1}}(u^k - A^Tz^{k+1}), \\ u^{k+1} &= u^k - A^Tz^{k+1} - y^{k+1}. \end{aligned} \tag{2.19}$$

When $M_1 = \frac{1}{\tau}I, M_2 = \lambda I$, (2.19) reduces to Linearized ADMM, also known as Split Inexact Uzawa [243].

Furthermore, iPrePDHG in Algorithm 2.1 is equivalent to (2.19) with inexact subproblems, which we present in Algorithm 2.2.

Algorithm 2.2 Inexact Preconditioned ADMM

Input: f, g, A in (2.1), preconditioners M_1 and M_2,

initial vector (z_0, y_0, u_0), subproblem solver S for the z-subproblem in (2.19), number of inner loops p, number of outer iterations K.

Output: (z^K, y^K, u^K)

for $k \leftarrow 0, 1, ..., K-1$ **do**
  $z_0^{k+1} = z^k$;
  for $i \leftarrow 0, 1, ..., p-1$ **do**
    $z_{i+1}^{k+1} = S(z_i^{k+1}, y^k, u^k)$;
  end for
  $z^{k+1} = z_p^{k+1}$; ▷ approximate $\mathrm{Prox}_{g^*}^{M_2}(z^k + M_2^{-1}AM_1^{-1}(-A^Tz^k - y^k + u^k))$.
  $y^{k+1} = \mathrm{Prox}_{f^*}^{M_1^{-1}}(u^k - A^Tz^{k+1})$;
  $u^{k+1} = u^k - A^Tz^{k+1} - y^{k+1}$;
end for

Proposition 2.3.5. *Under Assumption 2.3.1 and the transforms $u^k = M_1x^k$, $y^{k+1} = u^k - A^Tz^k - u^{k+1}$, PrePDHG (2.7) is equivalent to (2.19), and iPrePDHG in Algorithm 2.1 is equivalent to Algorithm 2.2.*

Proof. Set $u^k = M_1x^k$, $y^{k+1} = u^k - A^Tz^k - u^{k+1}$. Then (2.4) and (2.7) yield

$$y^{k+1} = M_1x^k - A^Tz^k - M_1x^{k+1} = \mathrm{Prox}_{f^*}^{M_1^{-1}}(u^k - A^Tz^k),$$

and

$$u^{k+1} = u^k - A^Tz^k - y^{k+1},$$
$$z^{k+1} = \mathrm{Prox}_{g^*}^{M_2}(z^k + M_2^{-1}AM_1^{-1}(-A^Tz^k - y^{k+1} + u^{k+1})).$$

If the z-update is performed first, then we arrive at (2.19).

In iPrePDHG or Algorithm 2.1, we are solving the z-subproblem of PrePDHG (2.7) approximately to the bounded relative error in Definition 2.3.1. This is equivalent to doing the same to the z-subproblem of (2.19), which yields Algorithm 2.2. □

Let us define the following generalized augmented Lagrangian:

$$L(z,y,u) = g^*(z) + f^*(y) + \langle -A^T z - y, M_1^{-1} u\rangle + \frac{1}{2}\|A^T z + y\|^2_{M_1^{-1}}. \tag{2.20}$$

Inspired by [231], we use (2.20) as the Lyapunov function to establish convergence of Algorithm 2.2 and, equivalently, the convergence of Algorithm 2.1. To the best of our knowledge, this is a new proof technique for inexact PDHG and inexact ADMM.

We first establish subsequential convergence of iPrePDHG in Algorithm 2.1 under the following additional assumptions.

Assumption 2.3.2.

1. *$f(x)$ is μ_f-strongly convex.*
2. *$g^*(z) + f^*(-A^T z)$ is coercive, i.e., $\lim_{\|z\|\to\infty} g^*(z) + f^*(-A^T z) = \infty$.*

To establish convergence of iPrePDHG in Algorithm 2.1, we also need the following assumption.

Assumption 2.3.3. *$L(z,y,u)$ is a KL function.*

Assumption 2.3.3 is true when both $g^*(z)$ and $f^*(y)$ are semi-algebraic, or more generally, definable in an o-minimal structure (more details can be referred to Sec 2.2 of [10] and Sec 2.2 of [233] and the references therein).

Theorem 2.3.6. *Take Assumptions 2.3.1 and 2.3.2. Choose any preconditioners M_1, M_2 and inner iteration number p such that*

$$C_1 = \frac{1}{2}M_1^{-1} - \frac{\|M_1\|}{\mu_f^2} I_n \succ 0, \tag{2.21}$$

$$C_2 = M_2 - \frac{1}{2}AM_1^{-1}A^T - c(p)I_m \succ 0, \tag{2.22}$$

where $c(p)$ depends on the z-subproblem iterator S and M_2 (e.g., (2.13) and (2.18)). Define $L^k := L(z^k, y^k, u^k)$. Then, Algorithm 2.2 satisfies the following sufficient descent

and lower boundedness properties, respectively:

$$L^k - L^{k+1} \geq \|y^k - y^{k+1}\|^2_{C_1} + \|z^k - z^{k+1}\|^2_{C_2}, \tag{2.23}$$

$$L^k \geq g^*(z^\star) + f^*(-A^T z^\star) > -\infty. \tag{2.24}$$

Proof. Since the z-subproblem of Algorithm 2.2 is solved to the bounded relative error in Def. 2.3.1, we have

$$\mathbf{0} \in \partial g^*(z^{k+1}) + M_2(z^{k+1} - z^k - M_2^{-1} A M_1^{-1}(-A^T z^k - y^k + u^k)) + \varepsilon^{k+1}, \tag{2.25}$$

where ε^{k+1} satisfies (2.12):

$$\|\varepsilon^{k+1}\| \leq c(p)\|z^{k+1} - z^k\|. \tag{2.26}$$

The y and u updates produce

$$\mathbf{0} = \nabla f^*(y^{k+1}) + M_1^{-1}(y^{k+1} - u^k + A^T z^{k+1}) = \nabla f^*(y^{k+1}) - M_1^{-1} u^{k+1}, \tag{2.27}$$

$$u^{k+1} = u^k - A^T z^{k+1} - y^{k+1}. \tag{2.28}$$

In order to show (2.23), let us write

$$\begin{aligned} g^*(z^k) &\geq g^*(z^{k+1}) \\ &\quad + \langle M_2(z^k - z^{k+1}) + A M_1^{-1}(-A^T z^k - y^k + u^k) - \varepsilon^{k+1}, z^k - z^{k+1} \rangle, \\ f^*(y^k) &\geq f^*(y^{k+1}) + \langle M_1^{-1} u^{k+1}, y^k - y^{k+1} \rangle. \end{aligned}$$

Assembling these inequalities with (2.26) gives us

$$\begin{aligned}
L^k - L^{k+1} &\geq \|z^k - z^{k+1}\|^2_{M_2 - c(p)I_m} \\
&\quad + \langle AM_1^{-1}(-A^T z^k - y^k + u^k), z^k - z^{k+1}\rangle + \langle M_1^{-1}u^{k+1}, y^k - y^{k+1}\rangle \\
&\quad + \langle -A^T z^k - y^k, M_1^{-1}u^k\rangle - \langle A^T z^{k+1} - y^{k+1}, M_1^{-1}(u^k - A^T z^{k+1} - y^{k+1})\rangle \\
&\quad + \frac{1}{2}\|A^T z^k + y^k\|^2_{M_1^{-1}} - \frac{1}{2}\|A^T z^{k+1} + y^{k+1}\|^2_{M_1^{-1}} \\
&= \|z^k - z^{k+1}\|^2_{M_2 - c(p)I_m} \\
&\quad + \langle AM_1^{-1}(-A^T z^k - y^k), z^k - z^{k+1}\rangle + \langle M_1^{-1}u^{k+1}, y^k - y^{k+1}\rangle \qquad \text{(A)} \\
&\quad + \langle -y^k, M_1^{-1}u^k\rangle - \langle -y^{k+1}, M_1^{-1}u^k\rangle \qquad \text{(B)} \\
&\quad + \frac{1}{2}\|A^T z^k + y^k\|^2_{M_1^{-1}} - \frac{3}{2}\|A^T z^{k+1} + y^{k+1}\|^2_{M_1^{-1}},
\end{aligned}$$

where the terms in (A) and (B) simplify to

$$\langle AM_1^{-1}(-A^T z^k - y^k), z^k - z^{k+1}\rangle + \langle M_1^{-1}(-A^T z^{k+1} - y^{k+1}), y^k - y^{k+1}\rangle. \tag{2.29}$$

Apply the following cosine rule on the two inner products above:

$$\langle a - b, a - c\rangle_{M_1^{-1}} = \frac{1}{2}\|a - b\|^2_{M_1^{-1}} + \frac{1}{2}\|a - c\|^2_{M_1^{-1}} - \frac{1}{2}\|b - c\|_{M_1^{-1}}.$$

Set $a = A^T z^k, c = A^T z^{k+1}$, and $b = -y^k$ to obtain

$$\begin{aligned}
\langle AM_1^{-1}(-A^T z^k - y^k), z^k - z^{k+1}\rangle &= -\frac{1}{2}\|A^T z^k + y^k\|^2_{M_1^{-1}} - \frac{1}{2}\|A^T z^k - A^T z^{k+1}\|^2_{M_1^{-1}} \\
&\quad + \frac{1}{2}\|y^k + A^T z^{k+1}\|^2_{M_1^{-1}}.
\end{aligned} \tag{2.30}$$

Set $a = y^{k+1}, c = y^k$, and $b = -A^T z^{k+1}$ to obtain

$$\begin{aligned}
\langle M_1^{-1}(-A^T z^{k+1} - y^{k+1}), y^k - y^{k+1}\rangle &= \frac{1}{2}\|A^T z^{k+1} + y^{k+1}\|^2_{M_1^{-1}} + \frac{1}{2}\|y^k - y^{k+1}\|_{M_1^{-1}} \\
&\quad - \frac{1}{2}\|A^T z^{k+1} + y^k\|^2_{M_1^{-1}}.
\end{aligned} \tag{2.31}$$

Combining (2.29), (2.30), and (2.31) yields

$$\begin{aligned}
L^k - L^{k+1} &\geq \|z^k - z^{k+1}\|^2_{M_2 - \frac{1}{2}AM_1^{-1}A^T - c(p)I_m} + \|y^k - y^{k+1}\|^2_{\frac{1}{2}M_1^{-1}} \\
&\quad - \|A^T z^{k+1} + y^{k+1}\|^2_{M_1^{-1}}.
\end{aligned} \tag{2.32}$$

Since f is μ_f-strongly convex, we know that ∇f^* is $\frac{1}{\mu_f}-$Lipschitz continuous. Consequently,

$$\|A^T z^{k+1} + y^{k+1}\|^2_{M_1^{-1}} = \|u^k - u^{k+1}\|^2_{M_1^{-1}} \leq \frac{1}{\lambda_{\min}(M_1^{-1})}\|M_1^{-1}(u^k - u^{k+1})\|^2 \overset{(2.27)}{\leq} \frac{\|M_1\|}{\mu_f^2}\|y^k - y^{k+1}\|^2. \tag{2.33}$$

Combining (2.32) and (2.33) gives us (2.23).

Now, to show (2.24), we use (2.27) and smoothness of f^* to get

$$f^*(y^k) \geq f^*(-A^T z^k) + \langle M_1^{-1}u^k, y^k + A^T z^k\rangle - \frac{1}{2\mu_f}\|A^T z^k + y^k\|^2.$$

Hence, we arrive at

$$\begin{aligned} L^k &= g^*(z^k) + f^*(y^k) + \langle -A^T z^k - y^k, M_1^{-1}u^k\rangle + \frac{1}{2}\|A^T z^k + y^k\|^2_{M_1^{-1}} \\ &\geq g^*(z^k) + f^*(-A^T z^k) + \frac{1}{2}\|A^T z^k + y^k\|^2_{M_1^{-1}} - \frac{1}{2\mu_f}\|A^T z^k + y^k\|^2. \end{aligned} \tag{2.34}$$

Since $C_1 \succ 0$ if and only if $\mu_f > \sqrt{2}\|M_1\|$, (2.24) follows. □

Next, we provide a simple choice of M_1, M_2, and p that ensures the positive definiteness of C_1 and C_2 in Theorem 2.3.6.

Proposition 2.3.7. *In order to ensure (2.21) and (2.22), it suffices to set $M_1 = \frac{1}{\tau}I_n$ where $\tau < \frac{1}{\sqrt{2}}\mu_f$, $M_2 = \tau AA^T + \theta I_m$ with any $\theta > 0$, and p is large enough such that $c(p) < \theta$.*

Proof. Since $M_1 = \frac{1}{\tau}I_n$, it is evident that $C_1 \succ 0$ if and only if $\tau < \frac{1}{\sqrt{2}}\mu_f$. With $M_1 = \frac{1}{\tau}I_n$ and $M_2 = \tau AA^T + \theta I_m$, we have

$$C_2 = \frac{1}{2}\tau AA^T + (\theta - c(p))I_m,$$

since $c(p)$ decreases linearly in p, we know that there exists p_0 such that $C_2 \succ 0$ for any $p \geq p_0$. □

We are now ready to show that (x^k, z^k) in Algorithm 2.1 converges subsequentially to a primal-dual solution pair of (2.1) and (2.2).

Theorem 2.3.8. *Take Assumptions 2.3.1 and 2.3.2. Then, (x^k, z^k) in Algorithm 2.1 is bounded, and any cluster point of $\{x^k, z^k\}$ is a primal-dual solution pair of (2.1) and (2.2).*

Proof. According to Theorem 2.3.5, it is sufficient to show that $\{M_1^{-1}u^k, z^k\}$ is bounded, and its cluster points are primal-dual solution pairs of (2.1).

Since L^k is nonincreasing, (2.34) tells us that

$$g^*(z^k) + f^*(-A^T z^k) + \frac{1}{2}\|A^T z^k + y^k\|^2_{M_1^{-1}} \leq L^0 < +\infty.$$

Since $g^*(z)+f^*(-A^T z)$ is coercive, $\{z^k\}$ is bounded, and, by the boundedness of $\{A^T z^k + y^k\}$, $\{y^k\}$ is also bounded. Furthermore, (2.27) gives us

$$\|M_1^{-1}(u^k - u^0)\| \leq \frac{1}{\mu_f}\|y^k - y^0\|.$$

Therefore, $\{M_1^{-1}u^k\}$ is bounded, too.

Let (z^c, y^c, u^c) be a cluster point of $\{z^k, y^k, u^k\}$. We shall show (z^c, y^c, u^c) is a saddle point of $L(z, y, u)$, i.e.,

$$\mathbf{0} \in \partial L(z^c, y^c, u^c), \tag{2.35}$$

or equivalently,

$$\begin{aligned}
\mathbf{0} &\in \partial g^*(z^c) - AM_1^{-1}u^c, \\
\mathbf{0} &= \nabla f^*(y^c) - M_1^{-1}u^c, \\
\mathbf{0} &= A^T z^c + y^c,
\end{aligned}$$

which ensures $(M_1^{-1}u^c, z^c)$ to be a primal-dual solution pair of (2.1).

In order to show (2.35), we first notice that (2.20) gives

$$\begin{aligned}
\partial_x L(z^{k+1}, y^{k+1}, u^{k+1}) &= \partial g^*(z^{k+1}) - AM_1^{-1}u^{k+1} + AM_1^{-1}(A^T z^{k+1} + y^{k+1}),\\
\nabla_y L(z^{k+1}, y^{k+1}, u^{k+1}) &= \nabla f^*(y^{k+1}) - M_1^{-1}u^{k+1} + M_1^{-1}(A^T z^{k+1} + y^{k+1}),\\
\nabla_u L(z^{k+1}, y^{k+1}, u^{k+1}) &= M_1^{-1}(-A^T z^{k+1} - y^{k+1}).
\end{aligned}$$

Comparing these with the optimality conditions (2.25), (2.27), and (2.28), we have

$$d^{k+1} = (d_z^{k+1}, d_y^{k+1}, d_u^{k+1}) \in \partial L(z^{k+1}, y^{k+1}, u^{k+1}), \tag{2.36}$$

where

$$\begin{aligned}
d_z^{k+1} &= M_2(z^k - z^{k+1}) + 2AM_1^{-1}(u^k - u^{k+1}) - AM_1^{-1}(u^{k-1} - u^k) - \varepsilon^{k+1},\\
d_y^{k+1} &= M_1^{-1}(u^k - u^{k+1}),\\
d_u^{k+1} &= M_1^{-1}(u^{k+1} - u^k).
\end{aligned} \tag{2.37}$$

Since (2.23) and (2.24) imply $z^k - z^{k+1}, y^k - y^{k+1} \to \mathbf{0}$, (2.27) gives $u^k - u^{k+1} \to \mathbf{0}$. Combine these with (2.12), we have $d^k \to \mathbf{0}$.

Finally, let us take a subsequence $\{z^{k_s}, y^{k_s}, u^{k_s}\} \to (z^c, y^c, u^c)$. Since $d^{k_s} \to \mathbf{0}$ as $s \to +\infty$, [194, Def. 8.3] and [194, Prop. 8.12] yield (2.35), which tells us that $(M_1^{-1}u^c, z^c)$ is a primal-dual solution pair of (2.1). □

Following the axiomatic approach developed in [10] for decent algorithms on KL functions, we can show that the whole sequence (x^k, z^k) in Algorithm 2.1 converges to a primal-dual solution pair. This approach has also been applied in [34] for KL-based Lagrangian optimization.

Theorem 2.3.9. *Take Assumptions 2.3.1, 2.3.2, and 2.3.3. Then, $\{x^k, z^k\}$ in Algorithm 2.1 converges to a primal-dual solution pair of (2.1).*

Proof. By Theorem 2.3.8, we can take $\{z^{k_s}, y^{k_s}, u^{k_s}\} \to (z^c, y^c, u^c)$ as $s \to \infty$. Since L is a KL function, we can prove the convergence of $\{z^k, y^k, u^k\}$ to $\{z^c, y^c, u^c\}$ following

the approach developed in [10]. Specifically, let us first verify that conditions H1, H2, and H3 of [10] are satisfied for $v^k := (z^k, y^k, u^k)$ and $L(v^k)$.

First, (2.23) gives

$$L(v^{k+1})+\lambda_{\min}(C_1)\|y^k - y^{k+1}\|^2 + \lambda_{\min}(C_2)\|z^k - z^{k+1}\|^2 \leq L(v^k). \tag{2.38}$$

By (2.27) and the $\frac{1}{\mu_f}-$Lipschitz differentiability of f^*, we know that

$$\frac{1}{2}\|y^k - y^{k+1}\|^2 \geq \frac{\mu_f^2}{2}\|M_1^{-1}u^k - M_1^{-1}u^{k+1}\|^2. \tag{2.39}$$

Combine (2.38) with (2.39), we know that there exists $a > 0$ such that

$$L(v^{k+1}) + a\|v^{k+1} - v^k\|^2 \leq L(v^k).$$

which satisfies condition H1 of [10].

From (2.36) and (2.37), we know that $d^{k+1} \in \partial L(v^{k+1})$ satisfies

$$\|d^{k+1}\| \leq b\|v^{k+1} - v^k\|$$

for some $b > 0$, which satisfies condition H2 of [10].

Next, let us verify that condition H3 of [10] also holds true.

Recall that we have taken $\{z^{k_s}, y^{k_s}, u^{k_s}\} \to (z^c, y^c, u^c)$ as $s \to \infty$. Note that $L(z^{k_s}, y^{k_s}, u^{k_s})$ is monotonic nonincreasing and lower bounded due to Theorem 2.3.6, which implies the convergence of $L(z^{k_s}, y^{k_s}, u^{k_s})$. Since L is lower semicontinuous, we have

$$L(z^c, y^c, u^c) \leq \lim_{s\to\infty} L(z^{k_s}, y^{k_s}, u^{k_s}). \tag{2.40}$$

Since the only potentially discontinuous term in L is g^*, we have

$$\lim_{s\to\infty} L(z^{k_s}, y^{k_s}, u^{k_s}) - L(z^c, y^c, u^c) \leq \limsup_{s\to\infty} g^*(z^{k_s}) - g^*(z^c). \tag{2.41}$$

By (2.25), we know that

$$\begin{aligned} g^*(z^c) &\geq g^*(z^{k_s}) \\ &+ \langle M_2(z^{k_s-1} - z^{k_s}) + AM_1^{-1}(-A^T z^{k_s-1} - y^{k_s-1} + u^{k_s-1}) - \varepsilon^{k_s}, z^c - z^{k_s}\rangle, \end{aligned}$$

Then, by Theorem 2.3.6, we further get $z^{k_s-1} - z^{k_s} \to \mathbf{0}$. Since $z^{k_s} \to z^c$ and $\{z^k, y^k, u^k\}$ is bounded, we obtain

$$\limsup_{s\to\infty} g^*(z^{k_s}) - g^*(z^c) \leq 0.$$

Combining this with (2.40) and (2.41), we conclude that

$$\lim_{s\to\infty} L(z^{k_s}, y^{k_s}, u^{k_s}) = L(z^c, y^c, u^c),$$

which satisfies condition H3 of [10].

Finally, since the conditions H1, H2, and H3 are satisfied, we can follow the proof of Theorem 2.9 of [10] to establish the convergence of $v^k = (z^k, y^k, u^k)$ to (z^c, y^c, u^c), which is a critical point of $L(z, y, u)$. By (2.37), we further now that $\{M_1^{-1}u^k, z^k\}$ converges to a primal-dual solution pair of (2.1), which is exactly $\{x^k, z^k\}$ in Algorithm 2.1 according to Theorem 2.3.5.

□

2.4 Numerical experiments

In this section, we compare our iPrePDHG (Algorithm 2.1) with (original) PDHG (2.6) and diagonally-preconditioned PDHG (DP-PDHG) [179]. We consider four popular applications of PDHG: TV-L^1 denoising, graph cuts, estimation of earth mover's distance, and CT reconstruction.

For the preconditioners M_1 and M_2 in iPrePDHG, we choose $M_1 = \frac{1}{\tau} I_n$ and $M_2 = \tau AA^T + \theta I$ as suggested in Proposition 2.3.7, which corresponds to ADMM and M_2 is nearly optimal for small θ (see subsection 2.3.2). The number of inner loops p is taken from $\{1, 2, 3\}$. Although f may not be strongly convex in our experiments, we still observe significant speedups compared to other algorithms.

When we write these examples in the form of (2.1), the matrix A (or a part of A) is one of the following operators:

Case 1: 2D discrete gradient operator $D : \mathbb{R}^{M\times N} \to \mathbb{R}^{2M\times N}$:

For images of size $M \times N$ and grid stepsize h, we have

$$(Du)_{i,j} = \begin{pmatrix} (Du)^1_{i,j} \\ (Du)^2_{i,j} \end{pmatrix},$$

where

$$(Du)^1_{i,j} = \begin{cases} \frac{1}{h}(u_{i+1,j} - u_{i,j}) & \text{if } i < M, \\ 0 & \text{if } i = M, \end{cases}$$

$$(Du)^2_{i,j} = \begin{cases} \frac{1}{h}(u_{i,j+1} - u_{i,j}) & \text{if } j < N, \\ 0 & \text{if } j = N. \end{cases}$$

where $w \in (\mathbb{R}^+)^{2MN}$ is a weight vector.

Case 2: 2D discrete divergence operator: div: $\mathbb{R}^{2M\times N} \to \mathbb{R}^{M\times N}$ given by

$$\text{div}(p)_{i,j} = h(p^1_{i,j} - p^1_{i-1,j} + p^2_{i,j} - p^2_{i,j-1}),$$

where $p = (p^1, p^2)^T \in \mathbb{R}^{2M\times N}$, $p^1_{0,j} = p^1_{M,j} = 0$ and $p^2_{i,0} = p^2_{i,N} = 0$ for $i = 1, ..., M$, $j = 1, ..., N$.

To take advantages of the finite-difference structure of these operators, we let S be the iterator of cyclic proximal BCD in Algorithm 2.1. We split $\{1, 2, ...m\}$ into 2 blocks (for case 3) or 4 blocks (for cases 1 and 2), which are inspired by the popular red-black ordering [201] for solving sparse linear system.

According to Theorem 2.3.4, running finitely many epochs of cyclic proximal BCD gives us a bounded relative error in Def. 2.3.1. We expect that this solver brings faster overall convergence. Specifically, when g^* is linear (or equivalently, g is a δ function), the z-subproblem in PrePDHG reduces to a linear system with a structured sparse matrix AA^T. Therefore, Gradient Descent amounts to the Richardson method [188, 201], and cyclic proximal BCD is equivalent to the Gauss-Seidel method [96, 201]. The following

two claims tell us that S in Algorithm 2.1 has a closed form, so Algorithm 2.1 is easy to implement. Furthermore, each execution of S can use parallel computing.

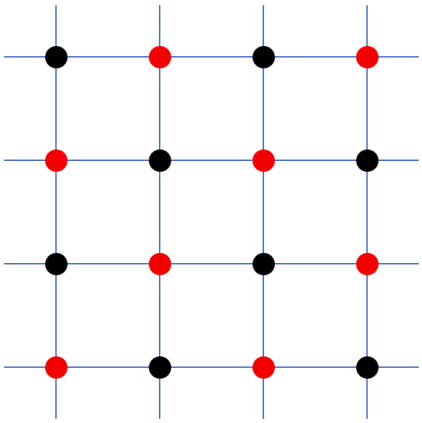

Figure 2.1: two-block ordering in Claim 2.4.1

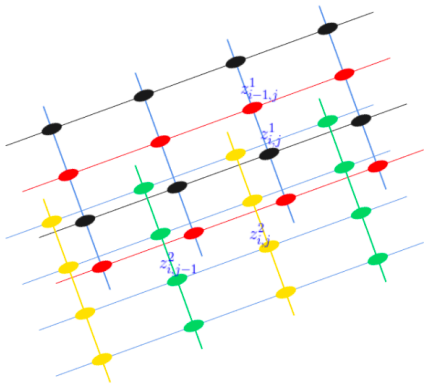

Figure 2.2: four-block ordering in Claim 2.4.2

Claim 2.4.1. *When $A = \text{div}$ (i.e. $A^T = -D$) and $M_2 = \tau AA^T$, for $z \in \mathbb{R}^{M\times N}$, we separate z into two block z_b, z_r where*

$$z_b := \{z_{i,j} \,|\, i+j \text{ is even}\},\ z_r := \{z_{i,j} \,|\, i+j \text{ is odd}\},$$

for $1 \le i \le M$, $1 \le j \le N$. If $g(z) = \Sigma_{i,j} g_{i,j}(z_{i,j})$ and $prox_{\gamma g^*_{i,j}}$ have closed-form solutions for all $1 \le i \le M$, $1 \le j \le N$ and $\gamma > 0$, then S as the iterator of cyclic proximal BCD in Algorithm 2.1 has a closed form and computing S is parallelizable.

Proof. As illustrated in Fig. 2.1, every black node is connected to its neighbor red nodes, so we can update all the coordinates corresponding to the black nodes in parallel, while those corresponding to the red nodes are fixed, and vice versa. See Appendix 2.C for a complete explanation. □

Claim 2.4.2. *When $A = D$ (i.e. $A^T = -\text{div}$) and $M_2 = \tau AA^T$, for $z = (z^1, z^2)^T \in$*

$\mathbb{R}^{2M\times N}$, *we separate* z *into four blocks* z_b, z_r, z_y *and* z_g, *where*

$$z_b = \{z^1_{i,j} \,|\, i \text{ is odd}\}, \quad z_r = \{z^1_{i,j} \,|\, i \text{ is even}\},$$
$$z_y = \{z^2_{i,j} \,|\, j \text{ is odd}\}, \quad z_g = \{z^2_{i,j} \,|\, j \text{ is even}\},$$

for $1 \le i \le M$, $1 \le j \le N$. *If* $g(z) = \Sigma_{i,j} g_{i,j}(z_{i,j})$ *and all* $prox_{\gamma g^*_{i,j}}$ *have closed-form solutions for all* $1 \le i \le M$, $1 \le j \le N$ *and* $\gamma > 0$, *then* S *as the iterator of cyclic proximal BCD in Algorithm 2.1 has a closed form and computing* S *is parallelizable.*

Proof. In Figure 2.2, the 4 blocks are in 4 different colors. The coordinates corresponding to nodes of the same color can be updated in parallel, while the rest are fixed. See Appendix 2.C for details. □

In Table 2.4.2, Table 2.4.1, Fig. 2.7, and Table 2.4.4, PDHG denotes original PDHG in (2.6) without any preconditioning; DP-PDHG denotes the diagonally-preconditioned PDHG in [179], PrePDHG denotes Preconditioned PDHG in (2.7) where the $(k+1)$th z-subproblem is solved until $\frac{\|z^k - z^{k+1}\|_2}{\max\{1, \|z^{k+1}\|_2\}} < 10^{-5}$ using the TFOCS [30] implementation of FISTA with restart; iPrePDHG (Inner: BCD) and iPrePDHG (S=FISTA) denote our iPrePDHG in Algorithm 2.1 with the iterator S being cyclic proximal BCD or FISTA with restart, respectly. All the experiments were performed on MATLAB R2018a on a MacBook Pro with a 2.5 GHz Intel i7 processor and 16GB of 2133MHz LPDDR3 memory.

A comparison between PDHG and DP-PDHG is presented in [179] on TV-L^1 denoising and graph cuts, and in [207] on CT reconstruction. A PDHG algorithm is proposed to estimate earth mover's distance (or optimal transport) in [127]. In order to provide a direct comparison, we use their problem formulations.

2.4.1 Graph cuts

The total-variation-based graph cut model involves minimizing a weighted TV energy:

$$\begin{aligned} \text{minimize} \quad & \|D_w u\|_1 + \langle u, \omega^u \rangle \\ \text{subject to} \quad & 0 \le u \le 1, \end{aligned}$$

where $w^u \in \mathbb{R}^{M \times N}$ is a vector of unary weights, $w^b \in \mathbb{R}^{2MN}$ is a vector of binary weights, and $D_w = \text{diag}(w^b)D$ for D being the 2D discrete gradient operator with $h = 1$.

To formulate this problem as (2.1), we take $f(u) = \langle u, w^u \rangle + \delta_{[0,1]}(u)$, $A = D$, and g as a weighted ℓ_1-norm:

$$g(z) = \sum_{i=1}^{2MN} (w^b)_i |z_i|.$$

In our experiment, the image has a size 660×720. We run all algorithms until $\delta^k := \frac{|\Phi^k - \Phi^\star|}{|\Phi^\star|} < 10^{-8}$, where Φ^k is the objective value at the kth iteration and Φ^* is the optimal objective value obtained by running CVX.

The best results of $\tau \in \{10, 1, 0.1, 0.01, 0.001\}$ and $p \in \{1, 2, 3\}$ are summarized in Table 2.4.1, where the step size of cyclic proximal BCD has been chosen as $\gamma = \frac{1}{\|M_2\|}$. We can see that our iPrePDHG (Inner: BCD) is the fastest. It is also worth mentioning that its number of outer iterations is close to that of PrePDHG, which solves z-subproblem much more accurately. In the last row of Table 2.4.1, we also take $M_2 = \tau D_w D_w^T + \theta I_m$ with $\theta > 0$ as suggest in Proposition 2.3.7, the performance is similar to that of $\theta = 0$. In practice, we recommend simply taking $\theta = 0$.

Method	Outer Iter	Runtime(s)	Parameters
PDHG	5529	140.5777	$\tau = 1, M_1 = \frac{1}{\tau} I_n, M_2 = \tau \|D_w\|^2 I_m$
DP-PDHG	3571	104.5392	$M_1 = \text{diag}(\Sigma_i \|D_{wi,j}\|)$, $M_2 = \text{diag}(\Sigma_j \|D_{wi,j}\|)$
PrePDHG (ADMM)	282	938.3787	$\tau = 10, M_1 = \frac{1}{\tau} I_n, M_2 = \tau D_w D_w^T$
iPrePDHG (Inner: BCD)	**411**	**14.9663**	$\tau = 10, M_1 = \frac{1}{\tau} I_n, M_2 = \tau D_w D_w^T, p = 2$
iPrePDHG (Inner: BCD)	**402**	**14.7687**	$\tau = 10, M_1 = \frac{1}{\tau} I_n, M_2 = \tau D_w D_w^T + \theta I_m, \theta = 0.1, p = 2$

Table 2.1: Graph cut test

Figure 2.3: Input image

Figure 2.4: Graph cut by iPrePDHG (Inner: BCD)

2.4.2 Total variation based image denoising

The following problem is known as the (discrete) TV-L^1 model for image denoising:

$$\text{minimize}_u \quad \Phi(u) = \|Du\|_1 + \lambda \|u - b\|_1,$$

where D is the 2D discrete gradient operator with $h = 1$, $b \in \mathbb{R}^{M \times N}$ is a noisy input image, and λ is a regularization parameter.

To formulate this problem as (2.1), we take $f(u) = \lambda\|u - b\|_1$, $g(z) = \|z\|_1$, and $A = D$.

In our experiment we input a 1024×1024 image with noise level 0.15 and set $\lambda = 1$; see Fig. 2.5. We run the algorithms until $\delta^k := \frac{|\Phi^k - \Phi^\star|}{|\Phi^\star|} < 10^{-6}$, where Φ^k is the objective value at kth iteration and Φ^* is the optimal objective value obtained by calling CVX [64, 107].

Observed performance is summarized in Table 2.4.2, where the best results for $\tau \in \{10, 1, 0.1, 0.01, 0.001\}$ and $p \in \{1, 2, 3\}$ are presented (Again, the step size of cyclic proximal BCD has been chosen as $\gamma = \frac{1}{\|M_2\|}$). Our iPrePDHG (Inner: BCD) is significantly faster than the other three algorithms.

When taking $\theta = 0.1$, we get nearly identical results. This is because $\theta > 0$ adds a proximal term $\frac{\theta}{2}\|z - z^k\|^2$ in the $z-$subproblem(see Equ. (2.8)), whose gradient at z^k is 0. Since $p = 1$ and cyclic proximal BCD is initialized exactly at z^k, we get the same iterates as that of $\theta = 0$. In practice, we recommend simply taking $\theta = 0$.

Remarkably, our algorithm uses fewer outer iterations than PrePDHG under the stopping criterion $\frac{\|z^k - z^{k+1}\|_2}{\max\{1, \|z^{k+1}\|_2\}} < 10^{-5}$, as this kind of stopping criteria may become looser as z^k is closer to $z^\star$. In this example, $\frac{\|z^k - z^{k+1}\|_2}{\max\{1, \|z^{k+1}\|_2\}} < 10^{-5}$ only requires 1 inner iteration of FISTA when Outer Iter $\geq$ 368, while as high as 228 inner iterations on average during the first 100 outer iterations. In comparison, our algorithm uses fewer outer iterations while each of them also costs less.

In addition, the diagonal preconditioner given in [179] appears to help very little when $A = D$. In fact, $M_1 = \mathrm{diag}(\Sigma_i |A_{i,j}|)$ will be $4I_n$ and $M_2 = \mathrm{diag}(\Sigma_j |A_{i,j}|)$ will be $2I_m$ if we ignore the Neumann boundary condition. Therefore, DP-PDHG performs even worse than PDHG.

Method	Outer Iter	Runtime(s)	Parameters
PDHG	2990	114.2576	$\tau = 0.01, M_1 = \frac{1}{\tau} I_n, M_2 = \tau \|D\|^2 I_m$
DP-PDHG	8856	329.7890	$M_1 = \text{diag}(\Sigma_i \|D_{i,j}\|), M_2 = \text{diag}(\Sigma_j \|D_{i,j}\|)$
PrePDHG (ADMM)	963	5706.2837	$\tau = 0.1, M_1 = \frac{1}{\tau} I_n, M_2 = \tau D D^T$
iPrePDHG (Inner: BCD)	**541**	**26.2704**	$\tau = 0.01, M_1 = \frac{1}{\tau} I_n, M_2 = \tau D D^T$, $p = 1$
iPrePDHG (Inner: BCD)	**541**	**26.2951**	$\tau = 0.01, M_1 = \frac{1}{\tau} I_n, M_2 = \tau D D^T + \theta I_m$, $p = 1, \theta = 0.1$

Table 2.2: TV-L^1 denoising test. PDHG is original PDHG. DP-PDHG uses diagonal preconditioning. PrePDHG uses non-diagonal preconditioning. iPrePDHG (Inner: BCD) is our algorithm that uses both non-diagonal preconditioning and an iterator S instead of solving the z-subproblem.

Figure 2.5: Noisy image

Figure 2.6: Denoising by iPrePDHG (Inner: BCD)

2.4.3 Earth mover's distance

Earth mover's distance is useful in image processing, computer vision, and statistics [122, 153, 174]. A recent method [127] to compute earth mover's distance is based on

$$\begin{aligned} &\text{minimize} && \|m\|_{1,2} \\ &\text{subject to} && \operatorname{div}(m) + \rho^1 - \rho^0 = 0, \end{aligned}$$

where $m \in \mathbb{R}^{2M\times N}$ is the sought flux vector on the $M \times N$ grid, and ρ^0, ρ^1 represents two mass distributions on the $M \times N$ grid. The setting in our experiment here is the same with that in [127], i.e. $M = N = 256$, $h = \frac{N-1}{4}$, and for ρ^0 and ρ^1 see Fig. 2.8.

To formulate this problem as (2.1), we take $f(m) = \|m\|_{1,2}$, $g(z) = \delta_{\{\rho^0-\rho^1\}}(z)$, and $A = \operatorname{div}$.

Since the iterates m^k may not satisfy the linear constraint, the objective $\Phi(m) = I_{\{m|\operatorname{div}(m)=\rho^0-\rho^1\}} + \|m\|_{1,2}$ is not comparable. Instead, we compare $\|m^k\|_{1,2}$ and the constraint violation until $k = 100000$ outer iterations in Fig. 2.4.3, where we set $\tau = 3 \times 10^{-6}$ as in [127], and $\sigma = \frac{1}{\tau\|\operatorname{div}\|^2}$. In Fig. 2.4.3, we can see that our iPrePDHG provides much lower constraint violation and much more faithful earth mover's distance $\|m\|_{1,2}$. Fig. 2.8 shows the solution obtained by our iPrePDHG (Inner: BCD), where m is the flux that moves the standing cat ρ^1 into the crouching cat ρ^0. For our iPrePDHG, when $M_2 = \tau\operatorname{div}\operatorname{div}^T + \theta I_m$, one has similar results for a small θ, the results are omitted. In practice, we recommend simply taking $\theta = 0$.

DP-PDHG and PrePDHG are extremely slow in this example. Similar to 2.4.2, when $A = \operatorname{div}$, the diagonal preconditioners proposed in [179] are approximately equivalent to fixed constant parameters $\tau = \frac{1}{2h}$, $\sigma = \frac{1}{4h}$ and they lead to extremely slow convergence. As for PrePDHG, it suffers from the high cost per outer iteration.

It is worth mentioning that unlike [127], the algorithms in our experiments are not parallelized. On the other hand, in our iPrePDHG (Inner: BCD), iterator S can be parallelized (which we did not implement). Therefore, one can expect a further speedup by a parallel implementation.

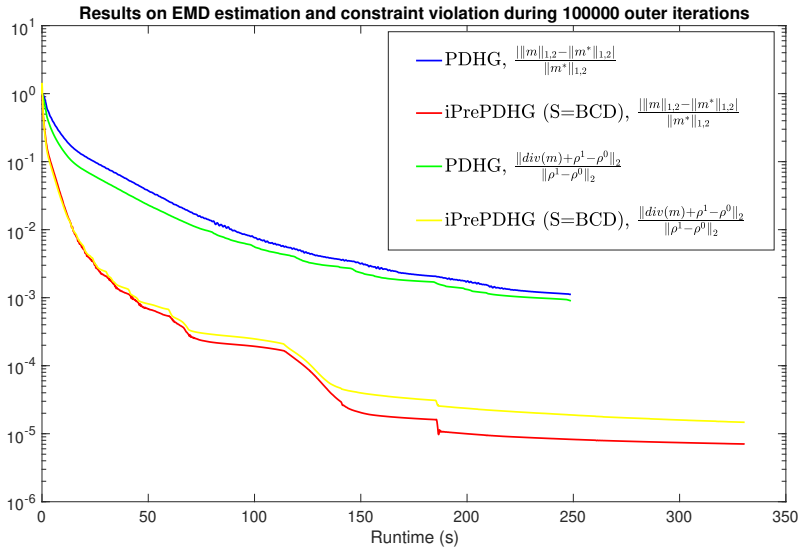

Figure 2.7: For PDHG, $\tau = 3 \times 10^{-6}$, $\sigma = \frac{1}{\tau\|\mathrm{div}\|^2}$; For iPrePDHG (Inner: BCD), $\tau = 3 \times 10^{-6}$, $M_1 = \tau^{-1} I_n$, $M_2 = \tau \mathrm{divdiv}^{\mathrm{T}}$, $\gamma = \frac{1}{\|M_2\|}$, and $p = 2$. $\|m^*|\|_{1,2}$ is obtained by calling CVX.

Figure 2.8: ρ^0, ρ^1 are the white standing cat and the black crouching cat, respectively. Both images are 256×256, and the earth mover's distance between ρ^0 and ρ^1 is 0.6718.

2.4.4 CT reconstruction

We test solving the following optimization problem for CT image reconstruction:

$$\text{minimize} \quad \Phi(u) = \tfrac{1}{2}\|Ru - b\|_2^2 + \lambda\|Du\|_1, \tag{2.42}$$

where $R \in \mathbb{R}^{13032\times 65536}$ is a system matrix for 2D fan-beam CT with a curved detector, $b = Ru_{\text{true}} \in \mathbb{R}^{13032}$ is a vector of line-integration values, and we want to reconstruct $u_{\text{true}} \in \mathbb{R}^{MN}$, where $M = N = 256$. D is the 2D discrete gradient operator with $h = 1$, and $\lambda = 1$ is a regularization parameter. By using the *fancurvedtomo* function from the AIR Tools II [111] package, we generate a test problem where the projection angles are $0°, 10°, \ldots, 350°$, and for all the other input parameters we use the default values.

Following [207], we formulate the problem (2.42) in the form of (2.1) by taking

$$g\begin{pmatrix} p \\ q \end{pmatrix} = \frac{1}{2}\|p - b\|_2^2 + \lambda\|q\|_1, \quad f(u) = 0, \quad A = \begin{pmatrix} R \\ D \end{pmatrix}, \tag{2.43}$$

By using this formulation, we avoids inverting the matrices R and D.

Since the block structure of AA^T is rather complicated, if we naively choose $M_1 = \frac{1}{\tau} I_n$ and $M_2 = \tau AA^T$ like in the previous three experiments, it becomes hard to find a fast subproblem solver for the z-subproblem. In Table 2.4.4, we report a TFOCS implementation of FISTA for solving the z-subproblem and the overall convergence is very slow.

Instead, we propose to choose

$$M_1 = \frac{2}{\tau} I_n, \quad M_2 = \begin{pmatrix} \tau \|R\|^2 I_{m-2n} & 0 \\ 0 & \tau DD^T \end{pmatrix} \tag{2.44}$$

or

$$M_1 = \operatorname{diag}(\Sigma_i |R_{i,j}|) + \frac{1}{\tau} I_n, \quad M_2 = \begin{pmatrix} \operatorname{diag}(\Sigma_j |R_{i,j}|) & 0 \\ 0 & \tau DD^T \end{pmatrix}. \tag{2.45}$$

These choices satisfy (2.9), and have simple block structures, a fixed epoch of S as cyclic proximal BCD iterators gives fast overall convergence. Note that (2.45) is a little slower but avoids the need of estimating $\|R\|$.

We summarize the numerical results in Table 2.4.4. All the algorithms are executed until $\delta^k := \frac{|\Phi^k - \Phi^\star|}{|\Phi^\star|} < 10^{-4}$, where Φ^k is the objective value at the kth iteration and Φ^* is the optimal objective value obtained by calling CVX. The best results of $\tau \in \{10, 1, 0.1, 0.01, 0.001\}$ and $p \in \{1, 2, 3\}$ are summarized in Table 2.4.4. As in the previous experiments, $\theta = 0.1$ gives similar performances for iPrePDHG(Inner: BCD). In practice, we recommend simply taking $\theta = 0$. For iPrePDHG (S=FISTA) with $M_2 = \tau AA^T$, the result for $p = 100$ is also reported (here we use the TFOCS implementation of FISTA).

2.5 Conclusion

We have developed an approach to accelerate PDHG and ADMM in this work. Our approach uses effective preconditioners to significantly reduce the number of iterations.

Method	Outer Iter	Runtime(s)	Parameters
PDHG	364366	3663.0348	$\tau = 0.001, M_1 = \frac{1}{\tau} I_n, M_2 = \tau \|A\|^2 I_m$
DP-PDHG	70783	713.9865	$M_1 = \text{diag}(\Sigma_i \|A_{i,j}\|)$, $M_2 = \text{diag}(\Sigma_j \|A_{i,j}\|)$
PrePDHG (ADMM)	-	$> 10^4$	$\tau = 0.01, M_1 = \frac{1}{\tau} I_n, M_2 = \tau A A^T$
iPrePDHG (Inner: FISTA)	-	$> 10^4$	$\tau = 0.001, M_1 = \frac{1}{\tau} I_n$, $M_2 = \tau A A^T, p = 1, 2, \text{or } 3$
iPrePDHG (Inner: FISTA)	-	$> 10^4$	$\tau = 0.01, M_1 = \frac{1}{\tau} I_n$, $M_2 = \tau A A^T, p = 100$
iPrePDHG (Inner: BCD)	**587**	**7.5365**	$\tau = 0.01, M_1 = \frac{2}{\tau} I_n, p = 2$, $M_2 = \begin{pmatrix} \tau \|R\|^2 I_{m-2n} & 0 \\ 0 & \tau D D^T \end{pmatrix}$
iPrePDHG (Inner: BCD)	**586**	**7.2112**	$\tau = 0.01, M_1 = \frac{2}{\tau} I_n, p = 2$, $M_2 = \begin{pmatrix} \tau \|R\|^2 I_{m-2n} & 0 \\ 0 & \tau D D^T + \theta I_{2n} \end{pmatrix}$
iPrePDHG (Inner: BCD)	**858**	**10.3517**	$\tau = 0.01, M_1 = \text{diag}(\Sigma_i \|R_{i,j}\|) + \frac{1}{\tau} I_n, p = 2$ $M_2 = \begin{pmatrix} \text{diag}(\Sigma_j \|R_{i,j}\|) & 0 \\ 0 & \tau D D^T \end{pmatrix}$
iPrePDHG (Inner: BCD)	**857**	**10.3123**	$\tau = 0.01, M_1 = \text{diag}(\Sigma_i \|R_{i,j}\|) + \frac{1}{\tau} I_n, p = 2$ $M_2 = \begin{pmatrix} \text{diag}(\Sigma_j \|R_{i,j}\|) & 0 \\ 0 & \tau D D^T + \theta I_{2n} \end{pmatrix}$

Table 2.3: CT reconstruction

In general, most effective preconditioners are non-diagonal and cause very difficult subproblems in PDHG and ADMM, so previous arts are restrictive with less effective diagonal preconditioners. However, we deal with those difficult subproblems by "solving" them highly inexactly, running just very few epochs of proximal BCD iterations. In all of our numerical tests, our algorithm needs relatively few outer iterations (due to effective preconditioners) and has the shortest total running time, achieving 4–95 times speedup over the state-of-the-art.

Theoretically, we show a fixed number of inner iterations suffice for global convergence though a new relative error condition. The number depends on various factors but is easy to choose in all of our numerical results.

There are still open questions left for us to address in the future: (a) Depending on problem structures, there are choices of preconditioners that are better than $M_1 = \frac{1}{\tau} I_n, M_2 = \tau A A^T$ (the ones that lead to ADMM if the subproblems are solved exactly). For example, in CT reconstruction, our choices of M_1 and M_2 have much faster overall convergence. (b) Is it possible to show Algorithm 2.1 converges even with S chosen as the iterator of faster accelerated solvers like APCG [131], NU_ACDM [3], and A2BCD [109]? (c) In general, how to accelerate a broader class of algorithms by integrating effective preconditioning and cheap inner loops while still ensuring global convergence?

2.A ADMM as a special case of PrePDHG

In this section we show that if we choose $M_1 = \frac{1}{\tau}$ and $M_2 = \tau A A^T$ in PrePDHG (2.7), then it is equivalent to ADMM on the primal problem (2.1).

By Theorem 1 of [237], we know that ADMM is primal-dual equivalent, in the sense that one can recover primal iterates from dual iterates and vice versa. Therefore, it suffices to show that $M_1 = \frac{1}{\tau}$ and $M_2 = \tau A A^T$ in PrePDHG (2.7) on the primal problem is equivalent to ADMM on the dual problem (2.2).

In Theorem 2.3.5 we have shown that, under an appropriate change of variables, PrePDHG on the primal is equivalent to applying (2.19) to the dual. As a result, we just need to demonstrate that the latter is exactly ADMM on the dual when $M_1 = \frac{1}{\tau} I_n$ and $M_2 = \tau A A^T$.

For the z-update in (2.19), we have

$$\begin{aligned}
z^{k+1} &= \arg\min_{z\in\mathbb{R}^m}\{g^*(z) - \tau\langle z - z^k, A(-A^T z^k - y^k + u^k)\rangle + \frac{\tau}{2}\|z - z^k\|^2_{AA^T}\} \\
&= \arg\min_{z\in\mathbb{R}^m}\{g^*(z) - \tau\langle z - z^k, A(-y^k + u^k)\rangle + \frac{\tau}{2}\|z\|^2_{AA^T}\} \\
&= \arg\min_{z\in\mathbb{R}^m}\{g^*(z) + \tau\langle z, A(y^k - u^k)\rangle + \frac{\tau}{2}\|A^T z\|^2\} \\
&= \arg\min_{z\in\mathbb{R}^m}\{g^*(z) + \tau\langle A^T z, -u^k\rangle + \frac{\tau}{2}\|A^T z + y^k\|^2\} \\
&= \arg\min_{z\in\mathbb{R}^m}\{g^*(z) + \tau\langle -A^T z - y^k, u^k\rangle + \frac{\tau}{2}\|A^T z + y^k\|^2\}. \qquad (2.46)
\end{aligned}$$

and for the y-update we have

$$\begin{aligned}
y^{k+1} &= \mathrm{Prox}^{M_1^{-1}}_{f^*}(u^k - A^T z^{k+1}) \\
&= \arg\min_{y\in\mathbb{R}^n}\{f^*(y) + \frac{\tau}{2}\|y - u^k + A^T z^{k+1}\|^2\} \\
&= \arg\min_{y\in\mathbb{R}^n}\{f^*(y) + \tau\langle -A^T z^{k+1} - y, u^k\rangle + \frac{\tau}{2}\|A^T z^{k+1} + y\|^2\}. \qquad (2.47)
\end{aligned}$$

Define $v^k = \tau u^k$, (2.46), (2.47), and the $u-$update in (2.19) become

$$\begin{aligned}
z^{k+1} &= \arg\min_{z\in\mathbb{R}^m}\{g^*(z) + \langle -A^T z - y^k, v^k\rangle + \frac{\tau}{2}\|A^T z + y^k\|^2\}, \\
y^{k+1} &= \arg\min_{y\in\mathbb{R}^n}\{f^*(y) + \langle -A^T z^{k+1} - y, v^k\rangle + \frac{\tau}{2}\|A^T z^{k+1} + y\|^2\}, \\
v^{k+1} &= v^k - \tau(A^T z^{k+1} + y^{k+1}),
\end{aligned}$$

which are ADMM iterations on the dual problem (2.2).

2.B Proof of Theorem 2.3.4: bounded relative error when S is the iterator of cyclic proximal BCD

The z-subproblem in (2.7) has the form

$$\min_{z\in\mathbb{R}^{\mathrm{m}}} h_1(z)+h_2(z),$$

where $h_1(z)=g^*(z)=\sum_{j=1}^{l} g_j^*(z_j)$, and $h_2(z)=\frac{1}{2}\|z-z^k-M_2^{-1}A(2x^{k+1}-x^k)\|_{M_2}^2$. And $z^{k+1}=z_p^{k+1}$ is given by

$$\begin{aligned} z_0^{k+1} &= z^k,\\ z_{i+1}^{k+1} &= S(z_i^{k+1},x^{k+1},x^k), \quad i=0,1,...,p-1, \end{aligned}$$

Here, S is the iterator of cyclic proximal BCD. Define

$$\begin{aligned} T(z) &= \mathrm{Prox}_{\gamma h_1(z)}(z-\gamma\nabla h_2(z)),\\ B(z) &= \frac{1}{\gamma}(z-T(z)), \end{aligned}$$

and the jth coordinate operator of B:

$$B_j(z)=(0,...,(B(z))_j,...,0), \quad j=1,2,...,l.$$

Then, we have

$$z_{i+1}^{k+1}=S(z_i^{k+1},x^{k+1},x^k)=(I-\gamma B_l)(I-\gamma B_2)...(I-\gamma B_1)z_i^{k+1}.$$

By [18, Prop. 26.16(ii)], we know that $T(z)$ is a contraction with coefficient $\rho_o=\sqrt{1-\gamma(2\lambda_{\min}(M_2)-\gamma\lambda_{\max}^2(M_2))}$. We know that for $\forall z_1,z_2\in\mathbb{R}^{\mathrm{m}}$ and $\mu_0=\frac{1-\rho_o}{\gamma}$,

$$\begin{aligned} \langle B(z_1)-B(z_2),z_1-z_2\rangle &= \frac{1}{\gamma}\|z_1-z_2\|^2-\frac{1}{\gamma}\langle T(z_1)-T(z_2),z_1-z_2\rangle\\ &\geq \mu_0\|z_1-z_2\|^2, \end{aligned}$$

Let $z_\star^{k+1}=\arg\min_{z\in\mathbb{R}^{\mathrm{m}}}\{h_1(z)+h_2(z)\}$. For [56, Thm 3.5], we have

$$\|z_i^{k+1}-z_\star^{k+1}\|\leq\rho^i\|z_0^{k+1}-z_\star^{k+1}\|, \quad \forall i=1,2,...,p. \tag{2.48}$$

where $\rho = 1 - \frac{\gamma\mu_0^2}{2}$.

Let $y_j = (I - \gamma B_j)...(I - \gamma B_1)z_{p-1}^{k+1}$ for $j = 1, ..., l$ and $y_0 = z_{p-1}^{k+1}$. Note that $(z_p^{k+1})_j = (y_j)_j$ for $j = 1, 2, ..., l$, and the blocks of y_j satisfies

$$(y_j)_t = \begin{cases} \Big(\text{Prox}_{\gamma g^*}\Big(y_{j-1} - \gamma\nabla h_2(y_{j-1})\Big)\Big)_t, & \text{if } t = j \\ (y_{j-1})_t, & \text{otherwise.} \end{cases}$$

On the other hand, we have

$$\text{Prox}_{\gamma g^*}\Big(y_{j-1} - \gamma\nabla h_2(y_{j-1})\Big) = \arg\min_{y\in\mathbb{R}^m}\{g^*(y) + \frac{1}{2\gamma}\|y - y_{j-1} + \gamma\nabla h_2(y_{j-1})\|^2\}.$$

Since g^* and $\|\cdot\|^2$ are separable, we obtain

$$\mathbf{0} \in \partial g_j^*((y_j)_j) + \frac{1}{\gamma}\Big((y_j)_j - (y_{j-1})_j + \gamma\Big(\nabla h_2(y_{j-1})\Big)_j\Big), \quad \forall j = 1, 2, ..., l,$$

or equivalently,

$$\mathbf{0} \in \partial g_j^*((z_p^{k+1})_j) + \frac{1}{\gamma}\Big((z_p^{k+1})_j - (z_{p-1}^{k+1})_j + \gamma\Big(\nabla h_2(y_{j-1})\Big)_j\Big), \quad \forall j = 1, 2, ..., l.$$

Therefore,

$$\mathbf{0} \in \partial g^*(z_p^{k+1}) + \frac{1}{\gamma}\Big(z_p^{k+1} - z_{p-1}^{k+1} + \gamma\xi_p\Big), \quad \forall j = 1, 2, ..., l,$$

where $(\xi_p)_j = \Big(\nabla h_2(y_{j-1})\Big)_j$ for $j = 1, 2, ..., l$. Comparing this with (2.11), we obtain

$$\varepsilon^{k+1} = \xi_p - \nabla h_2(z_p^{k+1}) + \frac{1}{\gamma}(z_p^{k+1} - z_{p-1}^{k+1}).$$

Notice that the first $j-1$ blocks of y_{j-1} are the same with those of $y_l = z_p^{k+1}$, and the rest of the blocks are the same with those of $y_0 = z_{p-1}^{k+1}$, so we have

$$\begin{aligned} \|\varepsilon^{k+1}\| &\le \sum_{j=1}^{l}\lambda_{\max}(M_2)\|y_{j-1} - z_p^{k+1}\| + \frac{1}{\gamma}\|z_p^{k+1} - z_{p-1}^{k+1}\| \\ &\le l\lambda_{\max}(M_2)\|z_{p-1}^{k+1} - z_p^{k+1}\| + \frac{1}{\gamma}\|z_p^{k+1} - z_{p-1}^{k+1}\| \\ &\le (l\lambda_{\max}(M_2) + \frac{1}{\gamma})(\|z_p^{k+1} - z_\star^{k+1}\| + \|z_{p-1}^{k+1} - z_\star^{k+1}\|) \end{aligned}$$

Combine this with (2.48)

$$\|\varepsilon^{k+1}\| \leq (l\lambda_{\max}(M_2) + \frac{1}{\gamma})(\rho^p + \rho^{p-1})\|z_0^{k+1} - z_\star^{k+1}\|. \tag{2.49}$$

Combining

$$\begin{aligned}\|z^{k+1} - z^k\| &= \|z_p^{k+1} - z_0^{k+1}\| \\ &\geq \|z_0^{k+1} - z_\star^{k+1}\| - \|z_p^{k+1} - z_\star^{k+1}\| \\ &\geq (1-\rho^p)\|z_0^{k+1} - z_\star^{k+1}\|\end{aligned}$$

with (2.49), we obtain

$$\|\varepsilon^{k+1}\| \leq \frac{(l\lambda_{\max}(M_2) + \frac{1}{\gamma})(\rho^p + \rho^{p-1})}{1-\rho^p}\|z^{k+1} - z^k\|.$$

2.C Two-block ordering in Claim 2.4.1 and four-block ordering in Claim 2.4.2

According to (2.8), when $M_2 = \tau AA^T$, the z-subproblem of Algorithm 2.1 is

$$z^{k+1} = \arg\min_{z\in\mathbb{R}^m}\{g^*(z) - \langle z - z^k, A(2x^{k+1} - x^k)\rangle + \frac{\tau}{2}\|A^T(z - z^k)\|_2^2\}. \tag{2.50}$$

Let us prove Claim 2.4.1 first. In that claim, $A = \text{div} \in \mathbb{R}^{MN\times 2MN}$ and $z \in \mathbb{R}^{MN}$. Following the definition of the sets z_b and z_r, we separate the MN columns of $A^T = -D$ into two blocks L_b, L_r by associating them with z_b and z_r, respectively. Therefore, we have $A^T z = L_b z_b + L_r z_r$ for any $z \in \mathbb{R}^{MN}$.

By the red-black ordering in Fig. 2.1, different columns of L_b are orthogonal one another, so ${L_b}^T L_b$ is diagonal. Similarly, ${L_r}^T L_r$ is also diagonal.

Define $c^k = -A(2x^{k+1} - x^k)$, and let b be the set of black nodes and r the set of red nodes. We can rewrite (2.50) as

$$\begin{aligned}z^{k+1} = \arg\min_{z_b, z_r\in\mathbb{R}^{MN/2}}\{&g_b^*(z_b) + g_r^*(z_r) + \langle z_b, c_b^k\rangle + \langle z_r, c_r^k\rangle \\ &+\frac{\tau}{2}\|L_b(z_b - z_b^k) + L_r(z_r - z_r^k)\|_2^2\},\end{aligned} \tag{2.51}$$

where $g_b^*(z_b) = \sum_{(i,j)\in b} g_{i,j}^*(z_{i,j})$, $g_r^*(z_r) = \sum_{(i,j)\in r} g_{i,j}^*(z_{i,j})$, and c_b^k, c_r^k are the coordinates of c^k associated with z_b and z_r, respectively.

Applying cyclic proximal BCD to black and red blocks with stepsize γ yields

$$z_b^{k+\frac{t+1}{p}} = \text{Prox}_{\gamma g_b^*}\left(z_b^{k+\frac{t}{p}} - \gamma\left(c_b^k + \tau L_b^T L_b(z_b^{k+\frac{t}{p}} - z_b^k) + \tau L_b^T L_r(z_r^{k+\frac{t}{p}} - z_r^k)\right)\right), \quad (2.52)$$

$$z_r^{k+\frac{t+1}{p}} = \text{Prox}_{\gamma g_r^*}\left(z_r^{k+\frac{t}{p}} - \gamma\left(c_r^k + \tau L_r^T L_b(z_b^{k+\frac{t+1}{p}} - z_b^k) + \tau L_r^T L_r(z_r^{k+\frac{t}{p}} - z_r^k)\right)\right), \quad (2.53)$$

for $t = 0, 1, ..., p-1$, where p is the number of inner iterations in Algorithm 2.1.

Since $\text{Prox}_{\gamma g_b^*} = \sum_{(i,j)\in b} \text{Prox}_{\gamma g_{i,j}^*}$, $\text{Prox}_{\gamma g_r^*} = \sum_{(i,j)\in r} \text{Prox}_{\gamma g_{i,j}^*}$ and $\text{Prox}_{\gamma g_{(i,j)}^*}$ are closed-form, (2.52) and (2.53) have closed-form solutions. Furthermore, the updates within each block can be done in parallel.

The proof of Claim 2.4.2 is similar. When $A = D$, we separate the columns of A^T into four blocks L_b, L_r, L_y, L_g by associating them with z_b, z_r, z_y ,z_g, respectively. Therefore, we have $A^T z = L_b z_b + L_r z_r + L_y z_y + L_g z_g$ for all $z \in \mathbb{R}^{2MN}$. Similarly, by the block design in Fig. 2.2, cyclic proximal BCD iterations have closed-form solutions, and updates within each block can be executed in parallel.

CHAPTER 3

Inexact Preconditioning for SVRG and Katyusha X

3.1 Introduction

Empirical risk minimization is an important class of optimization problems that has many applications in machine learning, especially in the large-scale setting. In this chapter, we formulate it as the minimization of the following objective

$$F(x) = f(x) + \psi(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \psi(x), \tag{3.1}$$

where the finite sum $f(x)$ is strongly convex, each $f_i(x)$ in the finite sum is smooth[1] and can be nonconvex, and the regularizer $\psi(x)$ is proper, closed, and convex, but may be nonsmooth. A nonzero $\psi(x)$ is desirable in many applications, for example, ℓ_1-regularization that induces sparsity in the solution. Allowing f_i to be nonconvex is also necessary in some applications, e.g., shift-and-invert approach to solve PCA [200].

3.1.1 Related Work

To obtain a high quality approximate solution $\hat{x}$ of (3.1), stochastic variance reduction algorithms are a class of preferable choices in the large-scale setting where n is huge. If each f_i is $\sigma-$strongly convex and $L-$smooth, and $\psi = 0$, then SVRG [114], SAGA [72], SAG [197], SARAH [165], SDCA [203], SDCA without duality [202], and Finito/MISO [73, 150] can find such a $\hat{x}$ within $\mathcal{O}\Big((n + \frac{L}{\sigma})\ln(\frac{1}{\varepsilon})\Big)$ evaluations of component gradients ∇f_i, while vanilla gradient descent needs $\mathcal{O}(n\frac{L}{\sigma}\ln\frac{1}{\varepsilon})$ evaluations. Recently, SCSG

[1] A function f is said to be smooth if its gradient ∇f is Lipschitz continuous.

improves this complexity to $\mathcal{O}\Big((n \wedge \frac{L}{\sigma\varepsilon} + \frac{L}{\sigma}) \ln \frac{1}{\varepsilon}\Big)$[2]. When $\psi \neq 0$, many of these algorithms can be extended accordingly and the same gradient complexity is preserved [232, 72, 204]. Among these methods, SVRG has been a popular choice due to its low memory cost.

When the condition number $\frac{L}{\sigma}$ is large, the performances of these variance reduction methods may degenerate considerably. In view of this, there have been many schemes that incorporate second-order information into the variance reduction schemes. In [105], the problem data is first transformed by linear sketching in order to decrease the condition number, then SVRG is applied. However, the strategy is only proposed for ridge regression and it is unclear whether it can be applied to other problems.

A larger family of algorithms, called Stochastic Quasi-Newton (SQN) methods, apply to more general settings. The idea is to first sample one or a few Hessian-vector products, then perform a L-BFGS type update on the approximate Hessian inverse H_k [47, 155, 106], then H_k is applied to the SVRG-type stochastic gradient as a preconditioner. That is,

$$w_{t+1} = w_t - \eta H_k \tilde{\nabla}_t,$$

where $\tilde{\nabla}_t$ is a variance-reduced stochastic gradient.

Linear convergence is established and competitive numerical performances are observed for SQN methods. However, the theoretical linear rate depends on the condition number of the approximate Hessian, which again depends poorly on the condition number of the objective, so it is not clear whether they are faster than SVRG in general. Furthermore, they do not support nondifferentiable regularizers nonconvexity of individual f_i. Recently, the first issue is partially resolved in [130], where the algorithm is at least as fast as SVRG. To deal with the second issue, [230] applied a H_k-preconditioned proximal mapping of ψ after H_k is applied to the variance reduced stochastic gradient, but in order to evaluate this mapping efficiently, H_k is required to be of the symmetric

[2] $a \wedge b := \min\{a, b\}$.

rank-one update form $\tau I_d + uu^T$, where $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix and $u \in \mathbb{R}^{\mathrm{d}}$. However, H_k is still ill-conditioned with a conditioner number of order $\mathcal{O}(\frac{1}{\varepsilon})$, therefore only a gradient complexity of order $\mathcal{O}\Big((n+\kappa\frac{1}{\varepsilon})\ln(\frac{1}{\varepsilon})\Big)$ can be guaranteed.

Another way of exploiting second-order information is to cyclically calculate one individual Hessian $\nabla^2 f_i$ (or an approximation of it) [196, 154], linear and locally super-linear convergence are established. However, they require at least an $O(n)$ amount of memory to store the local variables, which will be substantial when n is large.

Aside from exploiting second-order information, it is also possible to apply Nesterov-type acceleration to SVRG. Recently, Katyusha [1] and Katyusha X [2] are developed in this spirit. Katyusha X also applies to the sum-of-nonconvex setting where each f_i can be nonconvex. There are also "Catalyst" accelerated methods [129], where a small amount of strong convexity $\frac{c}{2}\|x-y^k\|^2$ is added to the objective and is minimized inexactly at each step, then Nesterov acceleration is applied. However, Catalyst methods have an additional $\ln k$ factor in gradient complexity over Katyusha and Katyusha X.

3.1.2 Our Contributions

1. We propose to accelerate SVRG and Katyusha X by a *fixed* preconditioner, as opposed to time-varying preconditioners in SQN methods. And the subproblems are solved with *fixed* number of simple subroutines.

2. If the preconditioner captures the second order information of f, then there will be significant accelerations. With a good preconditioner M, when $\kappa_f \in (n^{\frac{1}{2}}, n^2d^{-2})$, Algorithm 3.1 and Algorithm 3.2 are $\mathcal{O}(\frac{n^{\frac{1}{2}}}{\kappa_f})$ and $\mathcal{O}(\sqrt{\frac{n^{\frac{1}{2}}}{\kappa_f}})$ times faster than SVRG and Katyusha X in terms of gradient complexity, respectively. When $\kappa_f > n^2d^{-2}$, these numbers become $\mathcal{O}(\frac{d}{\sqrt{n\kappa_f}})$ and $\mathcal{O}(\frac{d}{n^{\frac{3}{4}}})$. We also demonstrate these accelerations for Lasso and Logistic regression.

3. Our acceleration applies to the sum-of-nonconvex setting, where $f(x)$ in (3.1)

is strongly convex, but each individual f_i can be nonconvex. We also allow a nondifferentiable regularizer $\psi(x)$.

3.2 Preliminaries and Assumptions

In addition to the preliminaries introduced in Sec. 1.4, we also need the following in the chapter.

We use $\lceil \cdot \rceil$ to denote the ceiling function. For $r \in (0, 1]$, $N \sim$ **Geom**(r) denotes a random variable N that obeys the geometric distribution, i.e., $N = k$ with probability $(1-r)^k r$ for $k \in \mathbb{N}$. We have $\mathbb{E}[N] = \frac{1-p}{p}$.

Definition 3.2.1. *We say that $f : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ is L_f^M-smooth under $\|\cdot\|_M$, if it is differentiable and satisfies*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_f^M}{2}\|y - x\|_M^2, \forall x, y \in \mathbb{R}^{\mathrm{d}}.$$

Definition 3.2.2. *We say that f is σ_f-strongly convex, if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma_f}{2}\|y - x\|^2, \forall x, y \in \mathbb{R}^{\mathrm{d}}.$$

We say that f is σ_f^M-strongly convex under $\|\cdot\|_M$, if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma_f^M}{2}\|y - x\|_M^2, \forall x, y \in \mathbb{R}^{\mathrm{d}}.$$

L_f^M-smoothness under $\|\cdot\|_M$ is equivalent to $\|\nabla f_i(x) - \nabla f_i(y)\|_{M^{-1}} \leq L_f^M \|x - y\|_M$. Also, σ_f^M-strong convexity is equivalent to $\|\nabla f_i(x) - \nabla f_i(y)\|_{M^{-1}} \geq \sigma_f^M \|x - y\|_M$. Cf. Section 2 of [204].

Definition 3.2.3. *We define the condition number of f under $\|\cdot\|_M$ as $\kappa_f^M := \frac{L_f^M}{\sigma_f^M}$.*

When $M = I$, we have $\kappa_f^M = \kappa_f := \frac{L_f}{\sigma_f}$.

In this chapter, we will choose M such that $\kappa_f^M \ll \kappa$. For example, if $f(x) = \frac{1}{2}x^T Q x$ where $Q \succ 0$ is ill-conditioned, by choosing $M = Q$ we have

$$\|\nabla f(x) - \nabla f(y)\|_{M^{-1}} \equiv \|x - y\|_Q,$$

which tells us that $L_f^M = \sigma_f^M = 1$ and $\kappa_f^M = 1$, while $\kappa_f = \kappa(Q) \gg 1$. That is, under $Q-$metric, $f(x)$ has a much smaller condition number and can be minimized easily.

Definition 3.2.4. *For a proper closed convex function $\phi : \mathbb{R}^{\mathrm{d}} \to \mathbb{R} \cup \{+\infty\}$, its subdifferential at $x \in \mathbf{dom}\, f$ is written as*

$$\partial\phi(x) = \{v \in \mathbb{R}^{\mathrm{d}} \mid \phi(z) \geq \phi(x) + \langle v, z - x\rangle \ \forall z \in \mathbb{R}^{\mathrm{d}}\}.$$

Definition 3.2.5. *For a proper closed convex function $\phi : \mathbb{R}^{\mathrm{d}} \to \mathbb{R} \cup \{+\infty\}$, its $M-$preconditioned proximal mapping with step size $\eta > 0$ is defined by*

$$\mathrm{Prox}_{\eta\psi}^M(x) = \arg\min_{y \in \mathbb{R}^{\mathrm{d}}} \{\psi(y) + \frac{1}{2\eta}\|x - y\|_M^2\}.$$

When $M = I$, this reduces to the classical proximal mapping.

Finally, let us list the assumptions that will be effective throughout this chapter.

Assumption 3.2.1. *In the objective function (3.1),*

1. *Each $f_i(x)$ is L_f-smooth and L_f^M-smooth under $\|\cdot\|_M$.*
2. *$f(x)$ is σ_f-strongly convex, and σ_f^M-strongly convex under $\|\cdot\|_M$, where $\sigma_f > 0$ and $\sigma_f^M > 0$.*
3. *The regularization term $\psi(x)$ is proper closed convex and $\mathrm{Prox}_{\eta\psi}$ is easy to compute.*

Remark 3.2.1.

1. *In Assumption 3.2.1, we only require $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ to be strongly convex, while each $f_i(x)$ can be nonconvex.*
2. *Several common choices of regularizers have simple proximal mappings. For example, when $\psi(x) = \lambda\|\cdot\|_1$ with $\lambda > 0$, $\mathrm{Prox}_{\eta\psi}$ can be computed component wise as*

$$\mathrm{Prox}_{\eta\psi}(x) = sign(x)\max\{|x| - \eta\lambda, 0\}.$$

3.3 Proposed Algorithms

As discussed in Sec. 3.1, SVRG and Katyusha X suffer from ill-conditioning like other first order methods. In this section, we propose to accelerate them by applying inexact preconditioning. Let us illustrate the idea as follows,

1. We would like to apply a preconditioner $M \succ 0$ to the gradient descent step in SVRG. i.e.,

$$\begin{aligned} w_{t+1} &= \mathrm{Prox}_{\eta\psi}^{M}(w_t - \eta M^{-1}\tilde{\nabla}_t) \\ &= \arg\min_{y\in\mathbb{R}^d}\{\psi(y) + \frac{1}{2\eta}\|y - w_t\|_M^2 + \langle\tilde{\nabla}_t, y\rangle\}. \end{aligned} \tag{3.2}$$

   where $\tilde{\nabla}_t$ is a variance-reduced stochastic gradient. When $\psi = 0$ and this minimization is solved exactly, we have $w_{t+1} = w_t - \eta M^{-1}\tilde{\nabla}_t$, which is a preconditioned gradient update.

2. However, solving (3.2) exactly may be expensive and impractical. In fact it suffices to solve it *highly inexactly* by *fixed* number of simple subroutines.

We summarize the resulted algorithm in Algorithm 3.1 and call it Inexact Preconditioned(IP-) SVRG. Compared to SVRG, the only difference lies in line 7.

Algorithm 3.1 Inexact Preconditioned SVRG(iPreSVRG)

Input: $F(\cdot) = \psi(\cdot) + \frac{1}{n}\sum_{i=1}^{n} f_i(\cdot)$, initial vector x^0, step size $\eta > 0$, preconditioner $M \succ 0$, number of epochs K.

Output: vector x^K

1: **for** $k \leftarrow 0, ..., K-1$ **do**
2: $D^k \sim$ **Geom**$(\frac{1}{m})$;
3: $w_0 \leftarrow x^k$, $g \leftarrow \nabla f(x^k)$;
4: **for** $t \leftarrow 0, ..., D^k$ **do**
5: pick $i_t \in \{1, 2, ..., n\}$ uniformly at random;
6: $\tilde{\nabla}_t = g + \Big(\nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_0)\Big)$;
7: $w_{t+1} \approx \arg\min_{y \in \mathbb{R}^d} \{\psi(y) + \frac{1}{2\eta}\|y - w_t\|_M^2 + \langle \tilde{\nabla}_t, y\rangle\}$;
8: **end for**
9: $x^{k+1} \leftarrow w_{D+1}$;
10: **end for**

Remark 3.3.1. *1. In line 2, the epoch length D^k obeys a geometric distribution and $\mathbb{E}[m^k] = m - 1$, this is for the purpose of simplifying analysis (motivated by [121, 2]), in practice one can just set $D^k = m - 1$. In our experiments, this still brings significant accelerations.*

2. The choice of m affects the performance. Intuitively, a larger m means more gradient evaluations per epoch, but also more progress per epoch. Theoretically, we show that $m = \lceil \frac{n}{1+pd} \rceil$ gives faster convergence than SVRG, where p is the number of subroutines used in Line 7.

3. In line 6, one can also sample a batch of gradients instead of one. It is straightforward to generalize our convergence results in Sec. 3.4 to this setting.

4. If $M = I$, line 7 reduces to

$$w_{t+1} = \text{Prox}_{\eta\psi}(w_t - \eta\tilde{\nabla}_t),$$

and Algorithm 3.1 reduces to SVRG.

For $M \not\propto I$, line 7 contains an optimization problem that may not have a closed form solution:

$$\arg\min_{y\in\mathbb{R}^d}\{\psi(y) + \frac{1}{2\eta}\|y - w_t\|_M^2 + \langle\tilde{\nabla}_t, y\rangle\}. \tag{3.3}$$

To solve it inexactly, we propose to apply *fixed* number of iterations of some simple subroutines, which are initialized at w_t. This procedure is summarized in Procedure 3.1.

Procedure 3.1 Procedure for solving (3.3) inexactly

Input: Iterator S, iterator step size $\gamma > 0$, number of iterations $p \geq 1$, problem data $\eta > 0, w_t, M \succ 0, \tilde{\nabla}_t, \psi(\cdot)$.

Output: vector w_{t+1}

1: $w_{t+1}^0 \leftarrow w_t$;

2: **for** $i \leftarrow 0, ..., p-1$ **do**

3: $\quad w_{t+1}^{i+1} = S(w_{t+1}^i, \eta, M, \tilde{\nabla}_t, \psi)$;

4: **end for**

5: $w_{t+1} \leftarrow w_{t+1}^p$;

Remark 3.3.2. *In Procedure 3.1, there are many choices for the iterator S, for example, one can use proximal gradient, FISTA [29] (or equivalently, Nesterov acceleration [160]), and FISTA with restart [167]. Under these choices, line 3 is easy to compute. For example, when S is the proximal gradient step, line 3 of Procedure 3.1 becomes*

$$w_{t+1}^{i+1} = \text{Prox}_{\gamma\psi}(w_{t+1}^i - \frac{\gamma}{\eta}M(w_{t+1}^i - w_t) - \gamma\tilde{\nabla}_t).$$

Now, let us also apply the inexact preconditioning idea to Katyusha X (Algorithm 2 of [2]). Similar to Katyusha X, we first apply a momentum step, then one epoch of iPreSVRG (i.e., line $2 \sim 9$ of Algorithm 3.1).

Algorithm 3.2 Inexact Preconditioned Katyusha X(iPreKatX)

Input: $F(x) = \psi(x) + \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, initial vector x^0, step size $\eta > 0$, preconditioner $M \succ 0$, momentum weight $\tau \in (0, 1]$, number of epochs K.

Output: vector y^K

1: $y_{-1} = y_0 \leftarrow x_0$;

2: **for** $k \leftarrow 0, ..., K-1$ **do**

3: $\quad x_{k+1} \leftarrow \frac{\frac{3}{2}y_k + \frac{1}{2}x_k - (1-\tau)y_{k-1}}{1+\tau}$;

4: $\quad y_{k+1} \leftarrow$ Algorithm 3.1$^{1\text{ep}}(F, M, x_{k+1}, \eta)$;

5: **end for**

Remark 3.3.3. *1. When $\tau = \frac{1}{2}$, one can show that $x_{k+1} \equiv y_k$, and Algorithm 3.2 reduces to Algorithm 3.1.*

2. When $M = I$ and the proximal mapping is solved exactly, Algorithm 3.2 reduces to Katyusha X.

3. The convergence of Algorithm 3.2 is established when $\tau = \frac{1}{2}\sqrt{\frac{1}{2}m\eta\sigma_f^M}$. In practice, we found that many other choices of τ also work.

3.4 Main Theory

In this section, we proceed to establish the convergence of Algorithm 3.1 and Algorithm 3.2. The key idea is that when the preconditioned proximal gradient update in (3.3) is solved inexactly as in Procedure 3.1, the error can be bounded by $\|w_{t+1} - w_t\|_M$, under which we can still establish the overall convergence of Algorithm 3.1 and Algorithm 3.2. Combine this with the fixed number of simple subroutines in Procedure 3.1, we obtain a much lower gradient complexity when $\kappa_f > n^{\frac{1}{2}}$.

All the proofs in this section are deferred to the supplementary material.

First, Let us analyze the error in the optimality condition of (3.3) when it is solved

inexactly by FISTA with restart as in Procedure 3.1. Specifically,

Let $h_1(y) = \psi(y)$ and $h_2(y) = \frac{1}{2\eta}\|y - w_t\|_M^2 + \langle \tilde{\nabla}, y\rangle$, then the subproblem (3.3) can be written as

$$\min_y \Psi(y) = h_1(y) + h_2(y).$$

Therefore, FISTA with restart applied to (3.3) can be summarized in the following algorithm.

Algorithm 3.3 FISTA with restart for solving (3.3)

Input: Iterator S, iterator step size $\gamma > 0$, number of iterations $p \geq 1$, problem data $\eta > 0, w_t, h_1(y) = \psi(y)$ and $h_2(y) = \frac{1}{2\eta}\|y - w_t\|_M^2 + \langle \tilde{\nabla}, y\rangle$.

$w_{t+1}^{(0,0)} = u_{t+1}^{(0,1)} \leftarrow w_t, \theta_0 = 1$
for $i \leftarrow 0, ..., r-1$ **do**
  for $j \leftarrow 0, ..., p_0 - 1$ **do**
    $\theta_0 = 1;$
    $w_{t+1}^{(i,j+1)} = \mathrm{Prox}_{\gamma h_1}\left(u_{t+1}^{(i,j+1)} - \gamma \nabla h_2(u_{t+1}^{(i,j+1)})\right);$
    $\theta_{j+1} = \frac{1+\sqrt{1+4\theta_j^2}}{2};$
    $u_{t+1}^{(i,j+2)} = w_{t+1}^{(i,j+1)} + \frac{\theta_j - 1}{\theta_{j+1}}(w_{t+1}^{(i,j+1)} - w_{t+1}^{(i,j)});$
  end for
  $w_{t+1}^{(i+1,0)} = u_{t+1}^{(i+1,1)} \leftarrow w_{t+1}^{(i,p_0)}$
end for
$w_{t+1} \leftarrow w_{t+1}^{(r-1,p_0)};$

Lemma 3.4.1. *Take Assumption 3.2.1. Suppose in Procedure 3.1, we choose S as the iterator of FISTA with restart*[1] *every $p_0 = \lceil 2e\sqrt{\kappa(M)}\rceil$ steps, with step size $\gamma = \frac{\eta}{\lambda_{\max}(M)}$ and restart it $(r-1)$ times (that is, $p = rp_0$ iterations in total). Then, $w_{t+1} = w_{t+1}^{(r-1,p_0)}$ is an approximate solution to (3.3) that satisfies*

$$\mathbf{0} \in \partial\psi(w_{t+1}) + \frac{1}{\eta}M(w_{t+1} - w_t) + \tilde{\nabla}_t + M\varepsilon_{t+1}^p, \tag{3.4}$$

$$\|\varepsilon_{t+1}^p\|_M \leq \frac{c(p)}{\eta}\|w_{t+1} - w_t\|_M, \tag{3.5}$$

where

$$c(p) = 14\kappa(M)\frac{\tau^p}{1-\tau^p},$$

and

$$\tau = (\frac{4\kappa(M)}{p_0^2})^{\frac{1}{2p_0}} \leq \exp(-\frac{1}{2e\sqrt{\kappa(M)}+1}) < 1.$$

With Lemma 3.4.1, the overall convergences of Algorithm 3.1 and 3.2 can be established. The analysis is similar to that of [2].

Theorem 3.4.2. *Under Assumption 3.2.1, let* $x^* = \arg\min_x F(x)$, $64\kappa_f^M c^2(p) \leq 1$, $\eta \leq \frac{1}{2\sqrt{m}L_f^M}$, *and* $m \geq 4$. *Then the iPreSVRG in Algorithm 3.1 satisfies*

$$\mathbb{E}[F(x^k) - F(x^*)] \leq \mathcal{O}\Big((\frac{1}{1+\frac{1}{4}m\eta\sigma^M})^k\Big). \tag{3.6}$$

Theorem 3.4.3. *Under Assumption 3.2.1, let* $x^* = \arg\min_x F(x)$, $64\kappa_f^M c^2(p) \leq 1$, $\tau = \frac{1}{2}\sqrt{\frac{1}{2}m\eta\sigma_f^M}$, $\eta \leq \frac{1}{2\sqrt{m}L_f^M}$, *and* $m \geq 4$. *Then the iPreKatX in Algorithm 3.2 satisfies*

$$\mathbb{E}[F(x^k) - F(x^*)] \leq \mathcal{O}\Big((\frac{1}{1+\frac{1}{2}\sqrt{\frac{1}{2}m\eta\sigma^M}})^k\Big). \tag{3.7}$$

Remark 3.4.1. *When* $M = I$, *we have* $c(p) = 0$, *and Theorems 3.4.2 and 3.4.3 recovers the Theorems D.1 and 4.3 of [2].*

In Theorems 3.4.2 and 3.4.3, we need the number of simple subroutines p to be large enough such that $64\kappa_f^M c^2(p) \leq 1$, the following Lemma provides a sufficient condition for this.

Lemma 3.4.4. *If the subproblem iterator* S *in Procedure 3.1 is FISTA with restart every* $p_0 = \lceil 2e\sqrt{\kappa(M)}\rceil$ *steps, and with step size* $\gamma = \frac{\eta}{\lambda_{\max}(M)}$, *then, in order for*

[1]FISTA with restart can be replaced with any iterator with Q-linear convergence on the iterates. In our experiments, FISTA also works, and a simple choice of $p = 20$ is enough.

$64\kappa_f^M c^2(p) \leq 1$ *to hold, it suffices to choose*

$$p = (2e\sqrt{\kappa(M)} + 1) \ln \frac{\sqrt{\kappa_f^M}\kappa(M) + \sqrt{c_1}}{c_1} \tag{3.8}$$
$$= \mathcal{O}\left(\sqrt{\kappa(M)} \ln\left(\sqrt{\kappa_f^M}\kappa(M)\right)\right)$$

where $c_1 = \frac{1}{64*14^2}$.

With (3.6), (3.7), and (3.8), we can now calculate the gradient complexities of Algorithm 3.1 and Algorithm 3.2, but let us first do that for SVRG and Katyusha X.

In Assumption 3.2.1, we have assumed that $\text{Prox}_{\eta\psi}(\cdot)$ is cheap to evaluate, therefore, each epoch of SVRG needs $n + m$ gradient evaluations, which is also true for Katyusha X. As a result, the gradient complexity for SVRG and Katyusha X to reach $\varepsilon-$suboptimality are:

$$C_1(m, \varepsilon) = \mathcal{O}(\frac{n+m}{\ln(1 + \frac{1}{4}m\eta\sigma)} \ln \frac{1}{\varepsilon}), \tag{3.9}$$

$$C_2(m, \varepsilon) = \mathcal{O}(\frac{n+m}{\ln(1 + \frac{1}{2}\sqrt{\frac{1}{2}m\eta\sigma})} \ln \frac{1}{\varepsilon}). \tag{3.10}$$

For Algorithm 3.1 and Algorithm 3.2, each iteration in Procedure 3.1 is at most as expensive as d gradient computations[1] and is operated p times, therefore, one epoch of iPreSVRG/iPreKatX needs at most $n + (1 + pd)m$ gradient computations.

Consequently, we can write the the gradient complexity for Algorithm 3.1 and Algorithm 3.2 to reach $\varepsilon-$suboptimality as:

$$C_1'(m, \varepsilon) = \mathcal{O}(\frac{n + (1+pd)m}{\ln(1 + \frac{1}{4}m\eta\sigma^M)} \ln \frac{1}{\varepsilon}), \tag{3.11}$$

$$C_2'(m, \varepsilon) = \mathcal{O}(\frac{n + (1+pd)m}{\ln(1 + \frac{1}{2}\sqrt{\frac{1}{2}m\eta\sigma^M})} \ln \frac{1}{\varepsilon}). \tag{3.12}$$

Remark 3.4.2. *1. According to Lemma 3.4.4, when S is FISTA with restart, it suffices to choose p by (3.8).*

2. When the preconditioner M is chosen appropriately, the step size η in (3.11) and (3.12) can be much larger than that of (3.9) and (3.10).

Finally, we can compare $C_1(m,\varepsilon)$, $C_2(m,\varepsilon)$ with $C_1'(m,\varepsilon)$, $C_2'(m,\varepsilon)$, respectively. It turns out that there is a significant speedup when $\kappa > n^{\frac{1}{2}}$.

Theorem 3.4.5. *Take Assumption 3.2.1. Let the iterator S in Procedure 3.1 be FISTA with restart, and an appropriate preconditioner M is chosen such that κ_f and $\kappa(M)$ are of the same order, and κ_f^M is small compared to them, then*

1. *if $\kappa_f > n^{\frac{1}{2}}$ and $\kappa_f < n^2 d^{-2}$, then*

$$\frac{\min_{m\geq 1} C_1'(m,\varepsilon)}{\min_{m\geq 1} C_1(m,\varepsilon)} \leq \mathcal{O}\Big(\frac{n^{\frac{1}{2}}}{\kappa_f}\Big). \tag{3.13}$$

2. *if $\kappa_f > n^{\frac{1}{2}}$ and $\kappa_f > n^2 d^{-2}$, then*

$$\frac{\min_{m\geq 1} C_1'(m,\varepsilon)}{\min_{m\geq 1} C_1(m,\varepsilon)} \leq \mathcal{O}(\frac{d}{\sqrt{n\kappa_f}}). \tag{3.14}$$

Theorem 3.4.6. *Take Assumption 3.2.1. Let the iterator S in Procedure 3.1 be FISTA with restart, and an appropriate preconditioner M is chosen such that κ_f and $\kappa(M)$ are of the same order, and κ_f^M is small compared to them, then*

1. *if $\kappa_f > n^{\frac{1}{2}}$ and $\kappa_f < n^2 d^{-2}$, then*

$$\frac{\min_{m\geq 1} C_2'(m,\varepsilon)}{\min_{m\geq 1} C_2(m,\varepsilon)} \leq \mathcal{O}\Big(\sqrt{\frac{n^{\frac{1}{2}}}{\kappa_f}}\Big). \tag{3.15}$$

2. *If $\kappa_f > n^{\frac{1}{2}}$ and $\kappa_f > n^2 d^{-2}$, then*

$$\frac{\min_{m\geq 1} C_2'(m,\varepsilon)}{\min_{m\geq 1} C_2(m,\varepsilon)} \leq \mathcal{O}(\frac{d}{n^{\frac{3}{4}}}). \tag{3.16}$$

In Section 3.5, we provide practical choices of M for Lasso and Logistic regression.

[1]For each iteration of Procedure 3.1, the most expensive step is multiplying M to some vector, which is often cheaper than d gradient computations.

3.5 Experiments

To investigate the practical performance of Algorithms 3.1 and 3.2, we test on three problems: Lasso, logistic regression, and a synthetic sum-of-nonconvex problem. For the first two, each function in the finite sum is convex. To guarantee that the objective is strongly convex, a small ℓ_2-regularization is added to Lasso and logistic regression.

In the following, we compare SVRG, iPreSVRG, Katyusha X, and iPreKatX on four datasets from LIBSVM[1]: `w1a.t` (47272 samples, 300 features), `protein` (17766 samples, 357 features), `cod-rna.t` (271617 samples, 8 features), `australian` (690 samples, 14 features), and one synthetic dataset. The implementation settings are listed below,

1. We choose the epoch length $m = 100$ in all experiments, since we found that the choices $m \in \{\frac{n}{4}, \frac{n}{2}, n\}$ need more gradient evaluations.

2. For iPrePDHG and iPreKatX, we use FISTA as the subproblem iterator S. If the preconditioner M is diagonal, then the number of subroutines for solving the subproblem is $p = 1$, if not, then we set $p = 20$.

3. In all the experiments, we tune the step size η and momentum weight τ to their optimal.

4. All algorithms are initialized at $x^0 = \mathbf{0}$.

5. All algorithms are implemented in Matlab R2015b. To be fair, except for the subproblem routines for inexact preconditioning, the other parts of the code are identical in all algorithms. The experiments are conducted on a Windows system with Intel Core i7 2.6 GHz CPU. The code is available at:

   https://github.com/uclaopt/IPSVRG.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

3.5.1 Lasso

We formulate Lasso as

$$\underset{x\in\mathbb{R}^{\mathrm{d}}}{\text{minimize}} \ \frac{1}{2n}\sum_{i=1}^{n}(a_i^T x - b_i)^2 + \lambda_1\|x\|_1 + \lambda_2\|x\|_2^2, \tag{3.17}$$

where $a_i \in \mathbb{R}^d$ are feature vectors and $b_i \in \mathbb{R}$ are labels. Note that the first term is equivalent to $\frac{1}{2n}\|Ax-b\|^2$, where $A = (a_1, a_2, ..., a_n)^T \in \mathbb{R}^{n\times d}$ and $b = (b_1, b_2, \ldots, b_n) \in \mathbb{R}^n$.

For Lasso as in (3.17), we provide two choices of preconditioner M,

1. When d is small, we choose

$$M_1 = \frac{1}{n}A^T A,$$

this is the exact Hessian of the smooth part of the objective.

2. When d is large and $A^T A$ is diagonally dominant, we choose

$$M_2 = \frac{1}{n}\mathrm{diag}(A^T A) + \alpha I,$$

where $\alpha > 0$. In this case, the subproblem (3.3) can be solved exactly with $p = 1$ iteration.

Our numerical results are presented in the following figures. We didn't observe significant accelerations of Katyusha X over SVRG and iPreKatX over iPrePDHG, and we suspect the reason is that $m = 100$ and the optimal choices of step size η make $m\eta\sigma_f > 1$ or $m\eta\sigma_f^M > 1$, thus the complexity in (3.10) and (3.12) are not better than (3.9) and (3.11), respectively.

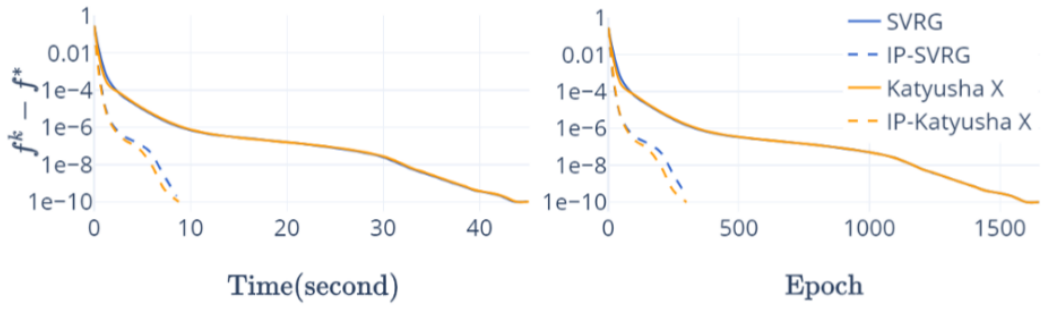

Figure 3.1: Lasso on `w1a.t`, $(n, d) = (47272, 300)$, $\lambda_1 = 10^{-3}, \lambda_2 = 10^{-8}$. For iPreSVRG and iPreKatX: $\eta_1 = 0.005$; For SVRG and Katyusha X: $\eta_2 = 0.08$; For Katyusha X and iPreKatX: $\tau = 0.45$, $M = M_2$ with $\alpha = 0.01$.

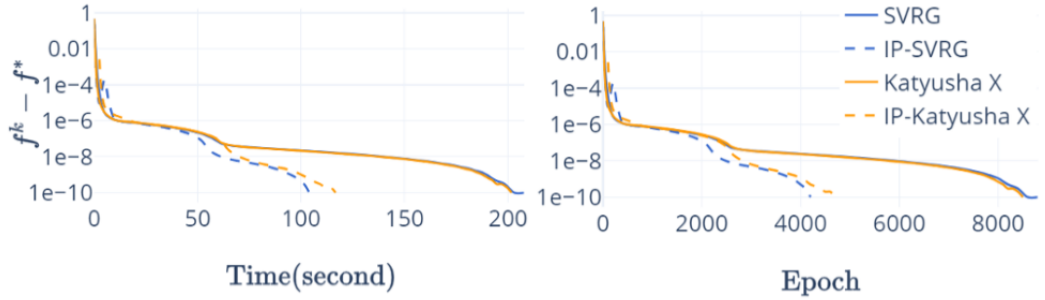

Figure 3.2: Lasso on `protein`, $(n, d) = (17766, 357)$, $\lambda_1 = 10^{-4}, \lambda_2 = 10^{-6}$, $\eta_1 = 0.008$, $\eta_2 = 0.2$, $\tau = 0.2$, $M = M_2$ with $\alpha = 0.008$.

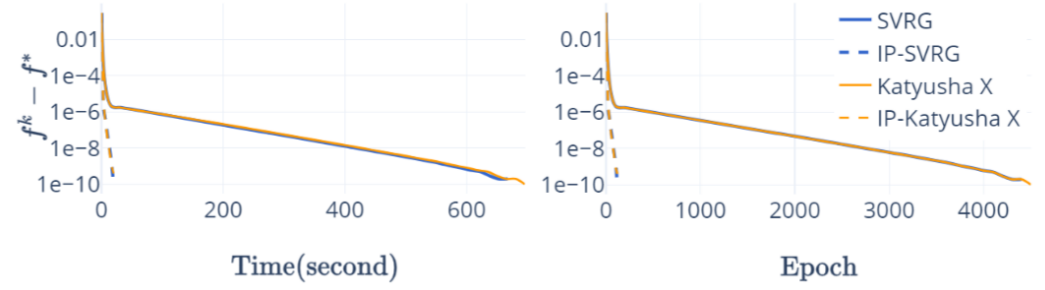

Figure 3.3: Lasso on `cod-rna.t`, $(n, d) = (271617, 8)$, $\lambda_1 = 10^{-2}, \lambda_2 = 1$, $\eta_1 = 1$, $\eta_2 = 5 \times 10^{-6}$, $\tau = 0.45$, $M = M_1$, subproblem iterator step size $\gamma = 3 \times 10^{-6}$.

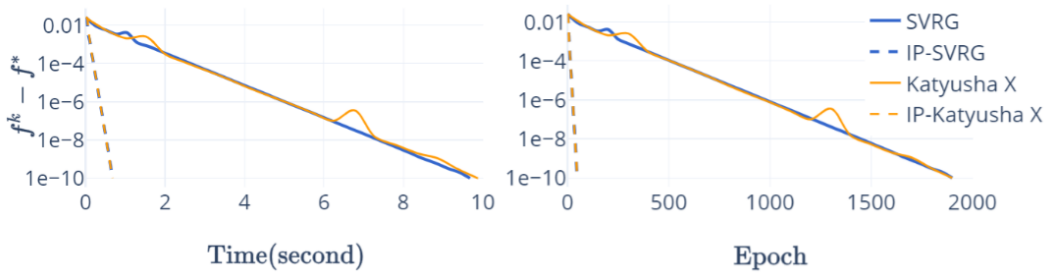

Figure 3.4: Lasso on `australian`, $(n, d) = (690, 14)$, $\lambda_1 = 2, \lambda_2 = 10^{-8}$, $\eta_1 = 0.01$, $\eta_2 = 8 \times 10^{-10}$, $\tau = 0.49$, $M = M_1$,$\gamma = 5 \times 10^{-10}$.

3.5.2 Logistic Regression

We formulate Logistic regression as

$$\underset{x\in\mathbb{R}^{\mathrm{d}}}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^{n} \ln\left(1+\exp(-b_i \cdot a_i^T x)\right) + \lambda_1\|x\|_1 + \lambda_2\|x\|_2^2, \tag{3.18}$$

where again $a_i \in \mathbb{R}^d$ are feature vectors and $b_i \in \mathbb{R}$ are labels.

For Logistic regression as in (3.18), the Hessian of the smooth part can be expressed as

$$H = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(-b_i a_i^T x)}{\left(1+\exp(-b_i a_i^T x)\right)^2} b_i^2 a_i a_i^T \preccurlyeq \frac{1}{4n} B^T B,$$

where $B = \mathrm{diag}(b)A = \mathrm{diag}(b)(a_1, a_2, ..., a_n)^T$. Inspired by this[1], we provide two choices of preconditioner M,

1. When d is small, we choose

$$M_1 = \frac{1}{4n} B^T B.$$

[1]Here is a heuristic justification: By Definition 3.2.1 we know that $L_f^M = 1$; Since $\frac{\exp(-b_i a_i^T x)}{\left(1+\exp(-b_i a_i^T x)\right)^2} \to 0$ only when x is unbounded, we know that if the iterates x^k of our algorithms are bounded, then $H(x^k) \succcurlyeq \frac{c}{n} B^T B$ for some $c > 0$, which gives $\sigma_f^M = 4c$ according to Definition 3.2.2. When c is not too small, one can expect $\kappa_f^M = \frac{1}{4c} \ll \kappa_f$.

2. When d is large and B^TB is diagonally dominant, we choose

$$M_2 = \frac{1}{4n}\mathrm{diag}(B^TB) + \alpha I,$$

where $\alpha > 0$. In this case, the subproblem (3.3) can be solved exactly with $p = 1$ iteration.

Our results are presented in the following figures, again, we didn't observe a significant acceleration of Katyusha X over SVRG and iPreKatX over iPrePDHG, due to the same reason mentioned in the last subsection.

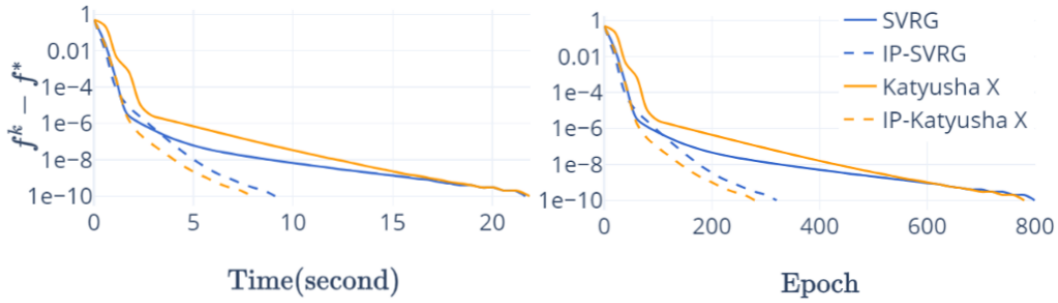

Figure 3.5: Logistic regression on `w1a.t`, $(n, d) = (47272, 300)$, $\lambda_1 = 5\times10^{-4}, \lambda_2 = 10^{-8}$, $\eta_1 = 0.06$, $\eta_2 = 4$, $\tau = 0.4$, $M = M_2$ with $\alpha = 0.005$.

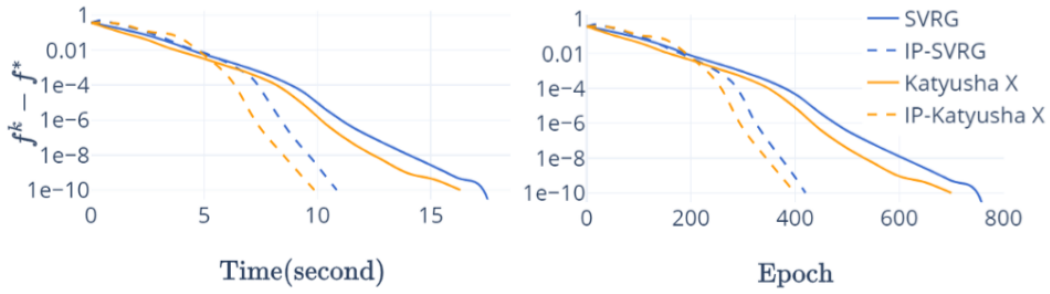

Figure 3.6: Logistic regression on `protein`, $(n, d) = (17766, 357)$, $\lambda_1 = 10^{-4}, \lambda_2 = 10^{-8}$, $\eta_1 = 1.5$, $\eta_2 = 10$, $\tau = 0.3$, $M = M_2$ with $\alpha = 0.05$.

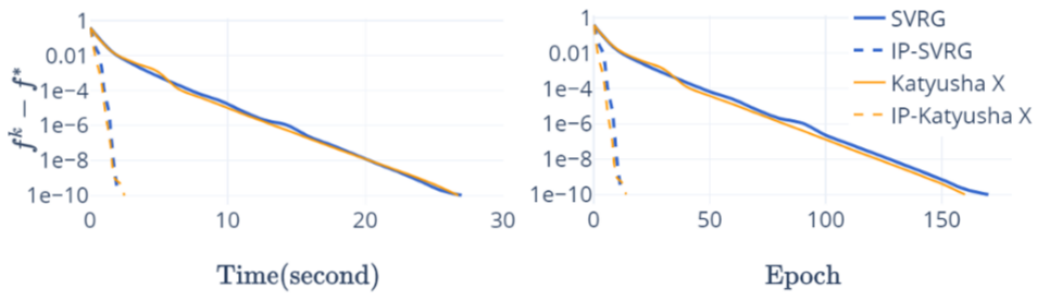

Figure 3.7: Logistic regression on `cod-rna.t`, $(n, d) = (271617, 8)$, $\lambda_1 = 0.1, \lambda_2 = 10^{-8}$, $\eta_1 = 1$, $\eta_2 = 3 \times 10^{-5}$, $\tau = 0.4$, $M = M_1$, $\gamma = 2 \times 10^{-5}$.

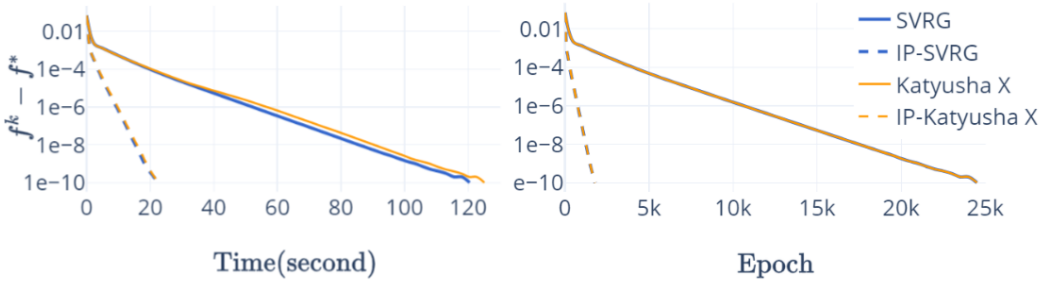

Figure 3.8: Logistic regression on `australian`, $(n, d) = (690, 14)$, $\lambda_1 = 0.5$, $\lambda_2 = 10^{-8}$, $\eta_1 = 1$, $\eta_2 = 10^{-6}$, $\tau = 0.2$, $M = M_1$, $\gamma = 2 \times 10^{-7}$.

3.5.3 Sum-of-nonconvex Example

Similar to [4], we generate a sum-of-nonconvex example by the following procedure:

We take n normalized random vector $a_i \in \mathbb{R}^{\mathrm{d}}$, and also d vectors of the form $g_i = (0, ...0, 5i, 0, ...0)$, where the nonzero element is at ith coordinate.

And the sum-of-nonconvex problem is given by

$$\underset{x \in \mathbb{R}^{\mathrm{d}}}{\text{minimize}} \ \frac{1}{2n} \sum_{i=1}^{n} x^T (c_i c_i^T + D_i) x + b^T x + \lambda_1 \|x\|_1, \tag{3.19}$$

where $n = 2000, d = 100$, and $\lambda_1 = 10^{-3}$.

$$c_i = \begin{cases} a_i + g_i & i = 1, 2, ..., d, \\ a_i & \text{otherwise.} \end{cases}$$

$$D_i = \begin{cases} -100I & i = 1, 2, ..., \frac{n}{2}, \\ 100I & \text{otherwise.} \end{cases}$$

Since the sum of D_i's is 0, they do not affect the condition number of the whole problem. However, it makes most of the first half of f_i to be highly nonconvex. Overall, the condition number of this problem is equal to that of $\sum_{i=1}^n c_i c_i^T$, which is approximately 10000 in our tested data.

Since $\sum_{i=1}^n c_i c_i^T$ is diagonally dominant, we select $M = \text{diag}(\frac{1}{n}\sum_{i=1}^n c_i c_i^T) + \alpha I$ as the preconditioner. Our algorithms also have significant acceleration in this sum-of-nonconvex setting.

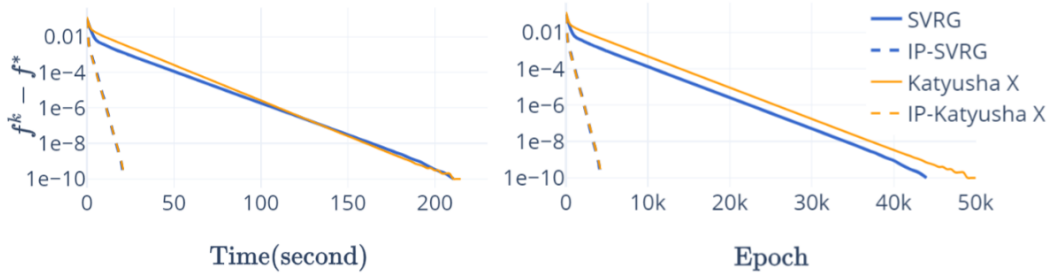

Figure 3.9: Sum-of-nonconvex on synthetic data. $\lambda_1 = 10^{-3}$, $\alpha = 15$. $\eta_1 = 0.015$, $\eta_2 = 10^{-4}$, $\tau = 0.45$.

3.6 Conclusions and Future Work

In this chapter, we accelerate SVRG and Katyusha X by inexact preconditioning, with an appropriate preconditioner, both can be provably accelerated in terms of iteration

complexity and gradient complexity. Our algorithms admits a nondifferentiable regularizer, as well as nonconvexity of individual functions. We confirm our theoretical results on Lasso, Logistic regression, and a sum-of-nonconvex example, where simple choices of preconditioners lead to significant accelerations.

There are still open questions left for us to address in the future: (a) Do we have theoretical guarantee when the subproblem iterator S is chosen as faster schemes such as APCG [131], NU_ACDM [3], and A2BCD [109]? (b) In general, how to choose a simple preconditioner that can greatly reduce the condition number of the problem? (c) Is it possible to apply this inexact preconditioning technique to other stochastic algorithms?

3.A Proof of Lemma 3.4.1

In this section, we prove the results on the error generated when solving the subproblem (3.3) inexactly by Procedure 3.1. Before proving Lemma 3.4.1, we will first prove a simpler case in Lemma 3.A.1, where the subproblem iterator S is the proximal gradient step.

Lemma 3.A.1. *Take Assumption 3.2.1. Suppose in Procedure 3.1, we choose S as the proximal gradient step with step size $\gamma = \eta\frac{\lambda_{\min}(M)}{\lambda^2_{\max}(M)}$, and is repeat it p times, where $p \geq 1$. Then, $w_{t+1} = w^p_{t+1}$ is an approximate solution to (3.3) that satisfies*

$$\mathbf{0} \in \partial\psi(w_{t+1}) + \frac{1}{\eta}M(w_{t+1} - w_t) + \tilde{\nabla}_t + M\varepsilon^p_{t+1}, \tag{3.20}$$

$$\|\varepsilon^p_{t+1}\|_M \leq \frac{c(p)}{\eta}\|w_{t+1} - w_t\|_M, \tag{3.21}$$

where

$$c(p) = (\kappa(M) + 1)\kappa(M)\frac{\tau^p + \tau^{p-1}}{1 - \tau^p},$$

and $\tau = \sqrt{1 - \kappa^{-2}(M)} < 1$.

Proof of Lemma 3.A.1. The optimization problem in (3.3) is of the form

$$\underset{y\in\mathbb{R}^{\mathrm{d}}}{\text{minimize}}\, h_1(y) + h_2(y), \tag{3.22}$$

for $h_1(y) = \psi(y)$ and $h_2(y) = \frac{1}{2\eta}\|y - w_t\|_M^2 + \langle\tilde{\nabla}, y\rangle$. With our choice of S as the proximal gradient descent step, the iterations in Procedure 3.1 are

$$\begin{aligned} w_{t+1}^0 &= w_t, \\ w_{t+1}^{i+1} &= \mathrm{Prox}_{\gamma h_1}\Big(w_{t+1}^i - \gamma\nabla h_2(w_{t+1}^i)\Big), \\ w_{t+1} &= w_{t+1}^p, \end{aligned}$$

where $i = 0, 1, ..., p-1$. From the definition of $\mathrm{Prox}_{\gamma h_1}$, we have

$$\mathbf{0} \in \partial h_1(w_{t+1}^p) + \nabla h_2(w_{t+1}^{p-1}) + \frac{1}{\gamma}(w_{t+1}^p - w_{t+1}^{p-1}).$$

Compare this with (3.20) gives

$$M\varepsilon_{t+1}^p = \frac{1}{\gamma}(w_{t+1}^p - w_{t+1}^{p-1}) + \nabla h_2(w_{t+1}^{p-1}) - \nabla h_2(w_{t+1}^p).$$

To bound the right hand side, let $w_{t+1}^\star$ be the solution of (3.22), $\alpha = \frac{\lambda_{\min}(M)}{\eta}$, and $\beta = \frac{\lambda_{\max}(M)}{\eta}$. Then $h_1(y)$ is convex and $h_2(y)$ is α-strongly convex and β-Lipschitz differentiable. Consequently, Prop. 26.16(ii) of [18] gives

$$\|w_{t+1}^i - w_{t+1}^\star\| \le \tau^i\|w_{t+1}^0 - w_{t+1}^\star\|, \quad \forall i = 0, 1, ..., p,$$

where $\tau = \sqrt{1 - \gamma(2\alpha - \gamma\beta^2)}$.

Let $a_i = \|w_{t+1}^i - w_{t+1}^\star\|$. Then, $a_i \le \tau^i a_0$. We can derive

$$\begin{aligned} \|M\varepsilon_{t+1}^p\| &\le (\frac{1}{\gamma} + \beta)\|w_{t+1}^p - w_{t+1}^{p-1}\| \\ &\le (\frac{1}{\gamma} + \beta)(a_p + a_{p-1}) \le (\frac{1}{\gamma} + \beta)(\tau^p + \tau^{p-1})a_0. \end{aligned}$$

On the other hand, we have

$$\|w_{t+1} - w_t\| \ge a_0 - a_p \ge (1 - \tau^p)a_0.$$

Combining these two equations yields

$$\|M\varepsilon_{t+1}^p\| \leq b(p)\|w_{t+1} - w_t\|, \tag{3.23}$$

where

$$b(p) = (\frac{1}{\gamma} + \frac{\lambda_{\max}(M)}{\eta})\frac{\tau^p + \tau^{p-1}}{1-\tau^p}. \tag{3.24}$$

Finally, let the eigenvalues of M be $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_d$, with orthonormal eigenvectors $v_1, v_2, ..., v_d$. Let ε_{t+1}^p and $w_{t+1} - w_t$ be decomposed by

$$\varepsilon_{t+1}^p = \sum_{i=1}^d \alpha_i v_i,$$
$$w_{t+1} - w_t = \sum_{i=1}^d \beta_i v_i.$$

then

$$\|\varepsilon_{t+1}^p\|_M = \sqrt{\sum_{i=1}^d \lambda_i \alpha_i^2} \leq \sqrt{\frac{1}{\lambda_{\min}(M)}\sum_{i=1}^d \lambda_i^2 \alpha_i^2} = \sqrt{\frac{1}{\lambda_{\min}(M)}}\|M\varepsilon_{t+1}^p\|,$$
$$\|w_{t+1} - w_t\| = \sqrt{\sum_{i=1}^d \beta_i^2} \leq \sqrt{\frac{1}{\lambda_{\min}(M)}\sum_{i=1}^d \lambda_i \beta_i^2} = \sqrt{\frac{1}{\lambda_{\min}(M)}}\|w_{t+1} - w_t\|_M.$$

Combine these two inequalities with (3.23), we arrive at

$$\|\varepsilon_{t+1}^p\|_M \leq c(p)\|w_{t+1} - w_t\|_M, \tag{3.25}$$

where

$$c(p) = \frac{1}{\lambda_{\min}(M)} b(p) = \frac{\frac{1}{\gamma} + \frac{\lambda_{\max}(M)}{\eta}}{\lambda_{\min}(M)} \frac{\tau^p + \tau^{p-1}}{1-\tau^p}.$$

□

Now, we are ready to prove Lemma 3.4.1, the techniques are similar to the proof of Lemma 3.A.1.

Proof of Lemma 3.4.1. We want to find $c(p)$ such that

$$\mathbf{0} \in \partial\psi(w_{t+1}) + \frac{1}{\eta}M(w_{t+1} - w_t) + \tilde{\nabla}_t + M\varepsilon_{t+1}^p, \tag{3.26}$$

$$\|\varepsilon_{t+1}^p\|_M \leq \frac{c(p)}{\eta}\|w_{t+1} - w_t\|_M, \tag{3.27}$$

Take $i = r-1$ and $j = p_0 - 1$, then the optimality condition of the problem in line 5 of Algorithm 3.3 is

$$\mathbf{0} \in \partial\psi(w_{t+1}^{(r-1,p_0)}) + \frac{1}{\gamma}(w_{t+1}^{(r-1,p_0)} - u_{t+1}^{(r-1,p_0)}) + \nabla h_2(u_{t+1}^{(r-1,p_0)}),$$

compare this with (3.26), we have

$$\begin{aligned} M\varepsilon_{t+1}^p =& \frac{1}{\gamma}(w_{t+1}^{(r-1,p_0)} - u_{t+1}^{(r-1,p_0)}) + \nabla h_2(u_{t+1}^{(r-1,p_0)}) - \frac{1}{\eta}M(w_{t+1} - w_t) - \tilde{\nabla}_t \\ =& \frac{1}{\gamma}(w_{t+1}^{(r-1,p_0)} - u_{t+1}^{(r-1,p_0)}) + \frac{1}{\eta}M(u_{t+1}^{(r-1,p_0)} - w_{t+1}) \end{aligned}$$

where

$$u_{t+1}^{(r-1,p_0)} = w_{t+1}^{(r-1,p_0-1)} + \frac{\theta_{p_0-2} - 1}{\theta_{p_0-1}}(w_{t+1}^{(r-1,p_0-1)} - w_{t+1}^{(r-1,p_0-2)}).$$

As a result,

$$\begin{aligned} \|M\varepsilon_{t+1}^p\| \leq& \|\frac{1}{\gamma}(w_{t+1}^{(r-1,p_0)} - u_{t+1}^{(r-1,p_0)})\| + \|\frac{1}{\eta}M(u_{t+1}^{(r-1,p_0)} - w_{t+1})\| \\ \leq& \|\frac{1}{\gamma}(w_{t+1}^{(r-1,p_0)} - w_{t+1}^{(r-1,p_0-1)}\| + \frac{1}{\gamma}\|\frac{\theta_{p_0-2}-1}{\theta_{p_0-1}}(w_{t+1}^{(r-1,p_0-1)} - w_{t+1}^{(r-1,p_0-2)})\| \\ &+ \|\frac{1}{\eta}M(w_{t+1}^{(r-1,p_0-1)} - w_{t+1})\| + \|\frac{1}{\eta}\frac{\theta_{p_0-2}-1}{\theta_{p_0-1}}M(w_{t+1}^{(r-1,p_0-1)} - w_{t+1}^{(r-1,p_0-2)})\|, \end{aligned} \tag{3.28}$$

Let the solution of (3.3) be $w_{t+1}^\star$. By Theorem 4.4 of [29], for any $0 \leq i \leq r-1$ and $0 \leq j \leq p_0$ we have

$$\Psi(w_{t+1}^{(i,j)}) - \Psi(w_{t+1}^\star) \leq \frac{2\lambda_{\max}(M)\|w_{t+1}^{(i,0)} - w_{t+1}^\star\|^2}{\eta j^2}.$$

On the other hand, the strong convexity of $\Psi = h_1 + h_2$ gives

$$\Psi(w_{t+1}^{(i,j)}) - \Psi(w_{t+1}^\star) \geq \frac{\lambda_{\min}(M)}{2\eta}\|w_{t+1}^{(i,j)} - w_{t+1}^\star\|^2.$$

Therefore,

$$\|w_{t+1}^{(i,j)} - w_{t+1}^\star\| \leq \sqrt{\frac{4\kappa(M)}{j^2}}\|w_{t+1}^{(i,0)} - w_{t+1}^\star\|. \tag{3.29}$$

Now, let us use (3.29) repeatedly to bound the right hand side of (3.28). For example, the first term can be bounded as

$$\begin{aligned}
&\|\frac{1}{\gamma}(w_{t+1}^{(r-1,p_0)} - w_{t+1}^{(r-1,p_0-1)}\| \\
\leq&\frac{1}{\gamma}\|w_{t+1}^{(r-1,p_0)} - w_{t+1}^{\star}\| + \frac{1}{\gamma}\|w_{t+1}^{(r-1,p_0-1)} - w_{t+1}^{\star}\| \\
\leq&\frac{1}{\gamma}(\frac{4\kappa(M)}{p_0^2})^{\frac{r}{2}}\|w_{t+1}^{(0,0)} - w_{t+1}^{\star}\| + \frac{1}{\gamma}(\frac{4\kappa(M)}{p_0^2})^{\frac{r-1}{2}}(\frac{4\kappa(M)}{(p_0-1)^2})^{\frac{1}{2}}\|w_{t+1}^{(0,0)} - w_{t+1}^{\star}\|.
\end{aligned}$$

Similarly, the rest of the terms can be bounded as follows,

$$\begin{aligned}
&\frac{1}{\gamma}\|\frac{\theta_{p_0-2}-1}{\theta_{p_0-1}}(w_{t+1}^{(r-1,p_0-1)} - w_{t+1}^{(r-1,p_0-2)})\| \\
\leq&\frac{1}{\gamma}(\frac{4\kappa(M)}{p_0^2})^{\frac{r-1}{2}}(\frac{4\kappa(M)}{(p_0-1)^2})^{\frac{1}{2}}\|w_{t+1}^{(0,0)} - w_{t+1}^{\star}\| + \frac{1}{\gamma}(\frac{4\kappa(M)}{p_0^2})^{\frac{r-1}{2}}(\frac{4\kappa(M)}{(p_0-2)^2})^{\frac{1}{2}}\|w_{t+1}^{(0,0)} - w_{t+1}^{\star}\|,
\end{aligned}$$

$$\begin{aligned}
&\|\frac{1}{\eta}M(w_{t+1}^{(r-1,p_0-1)} - w_{t+1})\| \\
\leq&\frac{\lambda_{\max}(M)}{\eta}(\frac{4\kappa(M)}{p_0^2})^{\frac{r-1}{2}}(\frac{4\kappa(M)}{(p_0-1)^2})^{\frac{1}{2}}\|w_{t+1}^{(0,0)} - w_{t+1}^{\star}\|, + \frac{\lambda_{\max}(M)}{\eta}(\frac{4\kappa(M)}{p_0^2})^{\frac{r}{2}}\|w_{t+1}^{(0,0)} - w_{t+1}^{\star}\|,
\end{aligned}$$

$$\begin{aligned}
&\|\frac{1}{\eta}\frac{\theta_{p_0-2}-1}{\theta_{p_0-1}}M(w_{t+1}^{(r-1,p_0-1)} - w_{t+1}^{(r-1,p_0-2)})\| \\
\leq&\frac{\lambda_{\max}(M)}{\eta}(\frac{4\kappa(M)}{p_0^2})^{\frac{r-1}{2}}(\frac{4\kappa(M)}{(p_0-1)^2})^{\frac{1}{2}}\|w_{t+1}^{(0,0)} - w_{t+1}^{\star}\| \\
&+ \frac{\lambda_{\max}(M)}{\eta}(\frac{4\kappa(M)}{p_0^2})^{\frac{r-1}{2}}(\frac{4\kappa(M)}{(p_0-2)^2})^{\frac{1}{2}}\|w_{t+1}^{(0,0)} - w_{t+1}^{\star}\|,
\end{aligned}$$

where in the first and third estimate we have used $\frac{\theta_{p_0-2}-1}{\theta_{p_0-1}} \leq \frac{\theta_{p_0-2}}{\theta_{p_0-1}} < 1$. On the other hand, we have

$$\begin{aligned}
\|w_{t+1} - w_t\| &= \|w_{t+1}^{(r-1,p_0)} - w_{t+1}^{(0,0)}\| \\
&\geq \|w_{t+1}^{(0,0)} - w_{t+1}^{\star}\| - \|w_{t+1}^{(r-1,p_0)} - w_{t+1}^{\star}\| \\
&\geq (1 - (\frac{4\kappa(M)}{p_0^2})^{\frac{r}{2}})\|w_{t+1}^{(0,0)} - w_{t+1}^{\star}\|.
\end{aligned}$$

As a result, taking $\gamma = \frac{\lambda_{\max}(M)}{\eta}$, $w_{t+1}^{(0,0)} = w_t$, $w_{t+1}^{(r-1,p_0)} = w_{t+1}$ and $\tau = (\frac{4\kappa(M)}{p_0^2})^{\frac{1}{2p_0}}$ yields

$$\|M\varepsilon_{t+1}^p\| \leq 2\frac{\lambda_{\max}(M)}{\eta}\frac{b(p)}{1-\tau^p}\|w_{t+1} - w_t\|,$$

where

$$b(p) = \tau^{p-p_0}\Big((\frac{4\kappa(M)}{(p_0-1)^2})^{\frac{1}{2}} + (\frac{4\kappa(M)}{(p_0-2)^2})^{\frac{1}{2}}\Big) + \tau^p + \tau^{p-p_0}(\frac{4\kappa(M)}{(p_0-1)^2})^{\frac{1}{2}}. \tag{3.30}$$

Similar to the end of proof of Lemma 3.A.1, we have

$$\|M\varepsilon_{t+1}^p\|_M \leq 2\frac{\kappa(M)}{\eta}\frac{b(p)}{1-\tau^p}\|w_{t+1}-w_t\|_M.$$

Now, let us choose p_0 such that $\tau = (\frac{4\kappa(M)}{p_0^2})^{\frac{1}{2p_0}}$ is minimized, a simple calculation yields

$$p_0^\star = 2e\sqrt{\kappa(M)}.$$

In order for p_0 to be an integer, we can take

$$p_0 = \lceil 2e\sqrt{\kappa(M)}\rceil,$$

then

$$\tau = (\frac{4\kappa(M)}{p_0^2})^{\frac{1}{2p_0}} \leq (\frac{1}{e^2})^{\frac{1}{2\lceil 2e\sqrt{\kappa(M)}\rceil}} \leq (\frac{1}{e^2})^{\frac{1}{2(2e\sqrt{\kappa(M)}+1)}} = \exp(-\frac{1}{2e\sqrt{\kappa(M)}+1}).$$

Finally, Let us show that $b(p)$ in (3.30) can be bounded by $7\tau^p$, and the desired bound (3.27) on $\|\varepsilon_{t+1}^p\|_M$ follows.

First, we have

$$\tau^{-p_0}(\frac{4\kappa(M)}{p_0-1})^{\frac{1}{2}} = (\frac{p_0}{p_0-1})^{\frac{1}{p_0}},$$

and

$$p_0 = \lceil 2e\sqrt{\kappa(M)}\rceil \geq \lceil 2e\rceil = 6.$$

On the other hand, a simple calculation shows that $(\frac{p_0}{p_0-1})^{\frac{1}{p_0}}$ is decreasing in p_0, therefore

$$\tau^{-p_0}(\frac{4\kappa(M)}{p_0-1})^{\frac{1}{2}} \leq (\frac{6}{5})^{\frac{1}{6}} < 2,$$

Similarly, one can show that

$$\tau^{-p_0}(\frac{4\kappa(M)}{p_0-2})^{\frac{1}{2}} \leq (\frac{6}{4})^{\frac{1}{6}} < 2.$$

Combining these two inequalities with (3.31) yields

$$b(p) \leq 7\tau^p.$$

□

3.B Proof of Theorem 3.4.2

In this section, we proceed to establish the convergence of inexact preconditioned SVRG as in Algorithm 3.1. The proof is similar to that of Theorem D.1 of [2].

Before proving Theorem 3.4.2, let us first prove several lemmas.

First, the inexact optimality condition (3.4) gives the following descent:

Lemma 3.B.1. *Under Assumption 3.2.1, suppose that (3.4) holds. Then, for any $u \in \mathbb{R}^{\mathrm{d}}$ we have*

$$\begin{aligned}\langle \tilde{\nabla}_t, w_t - u\rangle + \psi(w_{t+1}) - \psi(u) \leq & \langle \tilde{\nabla}_t, w_t - w_{t+1}\rangle + \frac{\|u-w_t\|_M^2}{2\eta} \\ & - \frac{1}{2\eta}\|u-w_{t+1}\|_M^2 - \frac{1}{2\eta}\|w_{t+1}-w_t\|_M^2 + \langle M\varepsilon_{t+1}^p, u - w_{t+1}\rangle.\end{aligned}$$

Proof. First, let us rewrite the left hand side as

$$\langle \tilde{\nabla}_t, w_t - u\rangle + \psi(w_{t+1}) - \psi(u) = \langle \tilde{\nabla}_t, w_t - w_{t+1}\rangle + \langle \tilde{\nabla}_t, w_{t+1} - u\rangle + \psi(w_{t+1}) - \psi(u).$$

By (3.4) and the definition of subdifferential we have

$$\psi(u) \geq \psi(w_{t+1}) - \langle \tilde{\nabla}_t + \frac{1}{\eta}M(w_{t+1}-w_t) + M\varepsilon_{t+1}^p, u - w_{t+1}\rangle.$$

Combining these two gives

$$\begin{aligned}\langle\tilde{\nabla}_t, w_t - u\rangle + \psi(w_{t+1}) - \psi(u) \leq & \langle\tilde{\nabla}_t, w_t - w_{t+1}\rangle + \langle\frac{1}{\eta}M(w_{t+1} - w_t) + M\varepsilon_{t+1}^p, u - w_{t+1}\rangle \\ = & \langle\tilde{\nabla}_t, w_t - w_{t+1}\rangle + \frac{\|u - w_t\|_M^2}{2\eta} \\ & - \frac{1}{2\eta}\|u - w_{t+1}\|_M^2 - \frac{1}{2\eta}\|w_{t+1} - w_t\|_M^2 + \langle M\varepsilon_{t+1}^p, u - w_{t+1}\rangle,\end{aligned}$$

where in the last equality we have applied

$$\langle a - b, c - a\rangle_M = -\frac{1}{2}\|a - b\|_M^2 - \frac{1}{2}\|a - c\|_M^2 + \frac{1}{2}\|b - c\|_M.$$

□

Based on lemma 3.B.1, we have

Lemma 3.B.2. *Under Assumption 3.2.1, if the iterator S in Procedure 3.1 is proximal gradient descent or FISTA with restart, then, for any $a > 0$, $\eta \leq \frac{1-2c(p)a}{2L_f^M}$, and $u \in \mathbb{R}^{\mathrm{d}}$ we have*

$$\begin{aligned}\mathbb{E}[F(w_{t+1}) - F(u)] \leq & \mathbb{E}[\eta\|\tilde{\nabla}_t - \nabla f(w_t)\|_{M^{-1}}^2 + \frac{1 - \eta\sigma_f^M}{2\eta}\|u - w_t\|_M^2 \\ & - (\frac{1}{2\eta} - \frac{c(p)}{2\eta a})\|u - w_{t+1}\|_M^2].\end{aligned}$$

Proof. We have

$$\begin{aligned}\mathbb{E}[F\left(w_{t+1}\right) - F(u)] = & \mathbb{E}[f\left(w_{t+1}\right) - f(u) + \psi\left(w_{t+1}\right) - \psi(u)] \\ \leq & \mathbb{E}[f\left(w_t\right) + \langle\nabla f\left(w_t\right), w_{t+1} - w_t\rangle + \frac{L_f^M}{2}\left\|w_t - w_{t+1}\right\|_M^2 - f(u) + \psi\left(w_{t+1}\right) - \psi(u)] \\ \leq & \mathbb{E}[\langle\nabla f\left(w_t\right), w_t - u\rangle - \frac{\sigma_f^M}{2}\left\|u - w_t\right\|_M^2 + \langle\nabla f\left(w_t\right), w_{t+1} - w_t\rangle \\ & + \frac{L_f^M}{2}\left\|w_t - w_{t+1}\right\|_M^2 + \psi\left(w_{t+1}\right) - \psi(u)] \\ = & \mathbb{E}[\langle\tilde{\nabla}_t, w_t - u\rangle - \frac{\sigma_f^M}{2}\left\|u - w_t\right\|_M^2 + \langle\nabla f\left(w_t\right), w_{t+1} - w_t\rangle \\ & + \frac{L_f^M}{2}\left\|w_t - w_{t+1}\right\|_M^2 + \psi\left(w_{t+1}\right) - \psi(u)], \qquad (3.31)\end{aligned}$$

where the first and second inequality are due to the smoothness and strong convexity under $\|\cdot\|_M$ in Assumption 3.2.1, respectively. The last equality is due to $\mathbb{E}[\tilde{\nabla}_t] = \nabla f(w_t)$.

On the other hand, recall that Lemma 3.B.1 gives

$$\langle \tilde{\nabla}_t, w_t - u\rangle + \psi(w_{t+1}) - \psi(u) \le \langle \tilde{\nabla}_t, w_t - w_{t+1}\rangle + \frac{\|u - w_t\|_M^2}{2\eta} - \frac{1}{2\eta}\|u - w_{t+1}\|_M^2 - \frac{1}{2\eta}\|w_{t+1} - w_t\|_M^2 + \langle M\varepsilon_{t+1}^p, u - w_{t+1}\rangle.$$

For the last term we can apply Cauchy-Schwartz as follows,

$$\langle M\varepsilon_{t+1}^p, u - w_{t+1}\rangle \le \|\varepsilon_{t+1}^p\|_M \|u - w_{t+1}\|_M,$$

from Lemma 3.A.1 and Lemma 3.4.1 we know that

$$\|\varepsilon_{t+1}^p\|_M \le \frac{c(p)}{\eta}\|w_{t+1} - w_t\|_M.$$

Therefore, by Young's inequality, we have for any $a > 0$ that

$$\langle M\varepsilon_{t+1}^p, u - w_{t+1}\rangle \le \frac{c(p)a}{2\eta}\|w_{t+1} - w_t\|_M^2 + \frac{c(p)}{2a\eta}\|u - w_{t+1}\|_M^2.$$

Applying this to Lemma 3.B.1 yields

$$\begin{aligned}\langle \tilde{\nabla}_t, w_t - u\rangle + \psi(w_{t+1}) - \psi(u) \le& \langle \tilde{\nabla}_t, w_t - w_{t+1}\rangle + \frac{\|u - w_t\|_M^2}{2\eta} - \frac{1}{2\eta}\|u - w_{t+1}\|_M^2 - \frac{1}{2\eta}\|w_{t+1} - w_t\|_M^2 \\ &+ \langle M\varepsilon_{t+1}^p, u - w_{t+1}\rangle \\ &\le \langle \tilde{\nabla}_t, w_t - w_{t+1}\rangle + \frac{\|u - w_t\|_M^2}{2\eta} - (\frac{1}{2\eta} - \frac{c(p)}{2a\eta})\|u - w_{t+1}\|_M^2 \\ &- (\frac{1}{2\eta} - \frac{c(p)a}{2\eta})\|w_{t+1} - w_t\|_M^2\end{aligned}$$

Applying this to (3.31), we arrive at

$$\begin{aligned}\mathbb{E}[F(w_{t+1}) - F(u)] \le& \mathbb{E}[\langle \tilde{\nabla}_t - \nabla f(w_t), w_t - w_{t+1}\rangle - \frac{1 - c(p)a - \eta L_f^M}{2\eta}\|w_t - w_{t+1}\|_M^2 \\ &+ \frac{1 - \eta\sigma_f^M}{2\eta}\|u - w_t\|_M^2 - (\frac{1}{2\eta} - \frac{c(p)}{2a\eta})\|u - w_{t+1}\|_M^2] \\ \le& \mathbb{E}[\frac{\eta}{2(1 - c(p)a - \eta L_f^M)}\|\tilde{\nabla}_t - \nabla f(w_t)\|_{M^{-1}}^2 \\ &+ \frac{1 - \eta\sigma_f^M}{2\eta}\|u - w_t\|_M^2 - (\frac{1}{2\eta} - \frac{c(p)}{2a\eta})\|u - w_{t+1}\|_M^2],\end{aligned}$$

where in the second inequality we have applied

$$\langle u_1, u_2\rangle = \langle M^{-\frac{1}{2}}u_1, M^{\frac{1}{2}}u_2\rangle \le \|u_1\|_{M^{-1}}\|u_2\|_M \le \frac{1}{2b}\|u_1\|^2_{M_1^{-1}} + \frac{b}{2}\|u_2\|^2_{M^{\frac{1}{2}}} \quad \text{for any } b>0.$$

Finally, since $\eta \le \frac{1-2c(p)a}{2L_f^M}$, we have $\frac{\eta}{2(1-c(p)a-\eta L_f^M)} \le \eta$, which gives the desired result.

□

Lemma 3.B.3. *Under Assumption 3.2.1, we have*

$$\mathbb{E}[\|\tilde{\nabla}_t - \nabla f(w_t)\|^2_{M^{-1}}] \le (L_f^M)^2\|w_0 - w_t\|^2_M.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}[\|\tilde{\nabla}_t - \nabla f(w_t)\|^2_{M^{-1}}] &= \mathbb{E}[\|\nabla f(w_0) + \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_0) - \nabla f(w_t)\|^2_{M^{-1}}] \\
&= \mathbb{E}[\|\Big(\nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_0)\Big) - \Big(\nabla f(w_t) - \nabla f(w_0)\Big)\|^2_{M^{-1}}] \\
&\le \mathbb{E}[\|\nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_0)\|^2_{M^{-1}}] \\
&\le (L_f^M)^2\|w_t - w_0\|^2_M,
\end{aligned}$$

where in the first inequality, we have applied $\mathbb{E}[\|\xi - \mathbb{E}\,\xi\|^2] = \mathbb{E}[\|\xi\|^2 - \|\,\mathbb{E}\,\xi\|^2$ with $\xi = M^{-\frac{1}{2}}\Big(\nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_0)\Big)$, and in the second inequality follows from Assumption 3.2.1. □

Lemma 3.B.4 ((Fact 2.3 of [2]))**.** *Let $C_1, C_2, \ldots$ be a sequence of numbers, and $N \sim$ **Geom**(p), then*

1. $\mathbb{E}_N[C_N - C_{N+1}] = \frac{p}{1-p}\mathbb{E}_N[C_0 - C_N]$, *and*
2. $\mathbb{E}_N[C_N] = (1-p)\mathbb{E}[C_{N+1}] + pC_0$.

Lemma 3.B.5. *Under Assumption 3.2.1, if $\eta \le \min\{\frac{1-2c(p)a}{2L_f^M}, \frac{1}{2\sqrt{m}L_f^M}\}$ and $m \ge 2$, then, for any $u \in \mathbb{R}^{\mathrm{d}}$ we have*

$$\begin{aligned}
\mathbb{E}[F(w_{D+1}) - F(u)] \le \mathbb{E}[&-\frac{1}{4m\eta}\|w_{D+1} - w_0\|^2_M + \frac{\langle w_0 - w_{D+1}, w_0 - u\rangle_M}{m\eta} \\
&- (\frac{\sigma_f^M}{4} - \frac{c(p)}{2a\eta})\|w_{D+1} - u\|^2_M].
\end{aligned}$$

Proof. By Lemmas 3.B.2 and 3.B.3, we know that

$$\mathbb{E}[F(w_{t+1}) - F(u)] \leq \mathbb{E}[\eta(L_f^M)^2\|w_0 - w_t\|_M^2 + \frac{1-\eta\sigma_f^M}{2\eta}\|u - w_t\|_M^2 - (\frac{1}{2\eta} - \frac{c(p)}{2\eta a})\|u - w_{t+1}\|_M^2].$$

Let $D \sim \mathbf{Geom}(\frac{1}{m})$ as in Algorithm 3.1 and take $t = D$, then

$$\begin{aligned}
\mathbb{E}[F(w_{D+1}) - F(u)] \leq & \mathbb{E}[\eta(L_f^M)^2\|w_0 - w_D\|_M^2 + \frac{1}{2\eta}\|u - w_D\|_M^2 \\
& - \frac{1}{2\eta}\|u - w_{D+1}\|_M^2 - \frac{\sigma_f^M}{2}\|u - w_D\|_M^2 + \frac{c(p)}{2\eta a}\|u - w_{D+1}\|_M^2] \\
= & \mathbb{E}[\eta(L_f^M)^2\|w_D - w_0\|_M^2 + \frac{\|u - w_0\|_M^2 - \|u - w_D\|_M^2}{2(m-1)\eta} \\
& - \frac{\sigma_f^M}{2}\|u - w_D\|_M^2 + \frac{c(p)}{2a\eta}\|u - w_{D+1}\|_M^2] \\
= & \mathbb{E}[\frac{m-1}{m}\eta(L_f^M)^2\|w_{D+1} - w_0\|_M^2 + \frac{\|u - w_0\|_M^2 - \|u - w_{D+1}\|_M^2}{2m\eta}] \\
& - \frac{\sigma_f^M}{2m}\|u - w_0\|_M^2 - \frac{\sigma_f^M(m-1)}{2m}\|u - w_{D+1}\|_M^2 + \frac{c(p)}{2a\eta}\|u - w_{D+1}\|_M^2] \\
\leq & \mathbb{E}[\eta(L_f^M)^2\|w_{D+1} - w_0\|_M^2 + \frac{\|u - w_0\|_M^2 - \|u - w_{D+1}\|_M^2}{2m\eta} \\
& - \frac{\sigma_f^M}{4}\|u - w_{D+1}\|_M^2 + \frac{c(p)}{2a\eta}\|u - w_{D+1}\|_M^2] \\
\leq & \mathbb{E}[-\frac{1}{4m\eta}\|w_0 - w_{D+1}\|_M^2 \\
& + \frac{\|u - w_0\|_M^2 - \|u - w_{D+1}\|_M^2 + \|w_0 - w_{D+1}\|_M^2}{2m\eta} \\
& - \frac{\sigma_f^M}{4}\|w_{D+1} - u\|_M^2 + \frac{c(p)}{2a\eta}\|u - w_{D+1}\|_M^2] \\
= & \mathbb{E}[-\frac{1}{4m\eta}\|w_{D+1} - w_0\|_M^2 + \frac{\langle w_0 - w_{D+1}, w_0 - u\rangle_M}{m\eta} \\
& - (\frac{\sigma_f^M}{4} - \frac{c(p)}{2a\eta})\|w_{D+1} - u\|_M^2],
\end{aligned}$$

where the first equality follows from the item 1 of Lemma 3.B.4 with $C_N = \|u - w_N\|_M^2$, the second inequality follows from item 2 with $C_N = \|w_d - w_0\|_M^2$, item 2 with $C_N = \|u - w_0\|_M^2 - \|u - w_N\|_M^2$, and item 1 with $C_N = \|u - w_D\|_M^2$, then third inequality makes use of $m \geq 2$ and the fourth inequality makes use of $\eta \leq \frac{1}{2\sqrt{m}L_f^M}$.

□

Now, let us proceed to prove Theorem 3.4.2. With Lemma 3.B.5, it can be proved in a similar way as Theorem 3 of [110].

Proof of Theorem 3.4.2. Without loss of generality, we can assume $x^\star = \arg\min_{x\in\mathbb{R}^{\mathrm{d}}} F(x) = \mathbf{0}$ and $F(x^*) = 0$.

According to Lemma 3.B.5, for any $u \in \mathbb{R}^d$, and $\eta \le \min\{\frac{1-2c(p)a}{2L_f^M}, \frac{1}{2\sqrt{m}L_f^M}\}$ we have

$$\begin{aligned}\mathbb{E}[F(x^{j+1}) - F(u)] \le &\mathbb{E}[-\frac{1}{4m\eta}\|x^{j+1} - x^j\|_M^2 \\ &+ \frac{\langle x^j - x^{j+1}, x^j - u\rangle_M}{m\eta} - (\frac{\sigma_f^M}{4} - \frac{c(p)}{2a\eta})\|x^{j+1} - u\|_M^2],\end{aligned}$$

or equivalently,

$$\begin{aligned}\mathbb{E}[F(x^{j+1}) - F(u)] \le &\mathbb{E}[\frac{1}{4m\eta}\|x^{j+1} - x^j\|_M^2 + \frac{1}{2m\eta}\|x^j - u\|_M^2 \\ &- \frac{1}{2m\eta}\|x^{j+1} - u\|_M^2 - (\frac{\sigma_f^M}{4} - \frac{c(p)}{2a\eta})\|x^{j+1} - u\|_M^2].\end{aligned}$$

In the following proof, we will omit $\mathbb{E}$.

Setting $u = x^* = 0$ and $u = x^j$ yields the following two inequalities:

$$F(x^{j+1}) \le \frac{1}{4m\eta}(\|x^{j+1} - x^j\|_M^2 + 2\|x^j\|_M^2) - \frac{1}{2m\eta}\Big(1 + \frac{1}{2}m\eta(\sigma_f^M - \frac{2c(p)}{a\eta})\Big)\|x^{j+1}\|_M^2, \tag{3.32}$$

$$F(x^{j+1}) - F(x^j) \le -\frac{1}{4m\eta}\Big(1 + m\eta(\sigma_f^M - \frac{2c(p)}{a\eta})\Big)\|x^{j+1} - x^j\|_M^2. \tag{3.33}$$

Define $\tau = \frac{1}{2}m\eta(\sigma_f^M - \frac{2c(p)}{a\eta})$, multiply $(1 + 2\tau)$ to (3.32), then add it to (3.33) yields

$$2(1+\tau)F(x^{j+1}) - F(x^j) \le \frac{1}{2m\eta}(1+2\tau)\Big(\|x^j\|_M^2 - (1+\tau)\|x^{j+1}\|_M^2\Big).$$

Multiplying both sides by $(1+\tau)^j$ gives

$$2(1+\tau)^{j+1}F(x^{j+1}) - (1+\tau)^jF(x^j) \le \frac{1}{2m\eta}(1+2\tau)\Big((1+\tau)^j\|x^j\|_M^2 - (1+\tau)^{j+1}\|x^{j+1}\|_M^2\Big).$$

Summing over $j = 0, 1, ..., k-1$, we have

$$(1+\tau)^k F(x^k) + \sum_{j=0}^{k-1}(1+\tau)^j F(x^j) - F(x^0) \leq \frac{1}{2m\eta}(1+2\tau)(\|x^0\|_M^2 - (1+\tau)^k\|x^k\|_M^2).$$

Since $F(x^j) \geq 0$, we have

$$F(x^k)(1+\tau)^k \leq F(x^0) + \frac{1}{2m\eta}(1+2\tau)\|x^0\|^2.$$

By the strong convexity of F, we have $F(x^0) \geq \frac{\sigma_f^M}{2}\|x^0\|_M^2$, therefore

$$F(x^k)(1+\tau)^k \leq F(x^0)(2 + \frac{1}{2\tau}). \tag{3.34}$$

Finally, recall that $a > 0$ can be chosen arbitrarily, so we can take

$$a = \frac{4c(p)}{\eta\sigma_f^M},$$

and

$$\eta \leq \min\{\frac{1-2c(p)a}{2L_f^M}, \frac{1}{2\sqrt{m}L_f^M}\} = \min\{\frac{1-\frac{8c^2(p)}{\eta\sigma_f^M}}{2L_f^M}, \frac{1}{2\sqrt{m}L_f^M}\}, \tag{3.35}$$

$$\tau = \frac{1}{2}m\eta(\sigma_f^M - \frac{2c(p)}{a\eta}) = \frac{1}{4}m\eta\sigma_f^M.$$

In order for the choice of η in (3.35) to be possible, we need

$$2L_f^M\eta^2 - \eta + 8\frac{c^2(p)}{\sigma_f^M} \leq 0 \tag{3.36}$$

to have one solution at least, which requires

$$64\kappa_f^M c^2(p) \leq 1,$$

under which $\eta = \frac{1}{4L_f^M}$ satisfy (3.36). As a result, $m \geq 4$ makes (3.35) into

$$\eta \leq \frac{1}{2\sqrt{m}L_f^M},$$

and the desired convergence result follows from (3.34). □

3.C Proof of Lemma 3.4.4

Proof. From Lemma 3.4.1, we know that

$$c(p) = 14\kappa(M)\frac{\tau^p}{1-\tau^p},$$

where

$$\tau \leq \exp(-\frac{1}{2e\sqrt{\kappa(M)}+1}).$$

Therefore, in order for $64\kappa_f^M c^2(p) \leq 1$, we need

$$\kappa_f^M \kappa^2(M)(\frac{\tau^p}{1-\tau^p})^2 \leq \frac{1}{64\times 14^2} = c_1,$$

which is equivalent to

$$\tau^p \leq \frac{c_1}{\sqrt{\kappa_f^M}\kappa(M)+\sqrt{c_1}}.$$

Thus, it suffices to require that

$$[\exp(-\frac{1}{2e\sqrt{\kappa(M)}+1})]^p \leq \frac{c}{\sqrt{\kappa_f^M}\kappa(M)+\sqrt{c_1}},$$

which gives

$$p \geq (2e\sqrt{\kappa(M)}+1)\ln\frac{\sqrt{\kappa_f^M}\kappa(M)+\sqrt{c_1}}{c_1}.$$

□

3.D Proof of Theorem 3.4.3

The proof of Theorem 3.4.3 is similar to that of Theorem 4.3 of [2], so we provide a proof sketch here and omit the details.

1. In [2], the proof of Theorem 4.3 is based on Lemma 3.3, here the proof of Theorem 3.4.3 is based on Lemma 3.B.5, which is an analog of Lemma of 3.3 in our settings.

2. Based on Lemma 3.B.5, the proof of Theorem 3.4.3 follows in nearly the same way as Theorem 4.3 of [2], the only difference is that one needs to replace σ by $\sigma_f^M - \frac{2c(p)}{a\eta}$.

3. By setting
$$a = \frac{4c(p)}{\eta\sigma_f^M},$$
and
$$64\kappa_f^M c^2(p) \leq 1$$
as in the proof of Theorem 3.4.2, the τ in Theorem 4.3 of [2] becomes $\frac{1}{2}m\eta\sigma_f^M$, and the convergence result of Theorem 3.4.3 follows.

3.E Proof of Theorems 3.4.5 and 3.4.6

Proof of Theorem 3.4.5. From Remark 3.4.1, we know that the gradient complexity of SVRG can be expressed as
$$C_1(m,\varepsilon) = \mathcal{O}(\frac{n+m}{\ln(1+\frac{1}{4}m\eta\sigma_f)}\ln\frac{1}{\varepsilon}).$$
Taking the largest possible step size $\eta = \frac{1}{2\sqrt{m}L_f}$ as in Theorem 3.4.2, we have
$$C_1(m,\varepsilon) = \mathcal{O}(\frac{n+m}{\ln(1+\frac{\sqrt{m}}{8\kappa_f})}\ln\frac{1}{\varepsilon}).$$
Let us first find the optimal $m = m^\star$ for SVRG, let
$$g(m) = \frac{n+m}{\ln(1+\frac{\sqrt{m}}{8\kappa_f})},$$
then
$$g'(m) = \frac{\ln(1+\frac{\sqrt{m}}{8\kappa_f}) - \frac{\frac{\sqrt{m}}{8\kappa_f}}{1+\frac{\sqrt{m}}{8\kappa_f}}\frac{n+m}{2m}}{\ln^2(1+z)}.$$
Taking derivative to the numerator gives
$$[\ln(1+\frac{\sqrt{m}}{8\kappa_f}) - \frac{\frac{\sqrt{m}}{8\kappa_f}}{1+\frac{\sqrt{m}}{8\kappa_f}}\frac{n+m}{2m}]' = (n+m)\frac{\frac{1}{32\kappa_f}m^{-\frac{3}{2}} + 2\frac{m^{-1}}{(16\kappa_f)^2}}{(1+\frac{\sqrt{m}}{8\kappa_f})^2} > 0,$$

Therefore, $m^\star$ is given by $g'(m) = 0$. Let $z = \frac{\sqrt{m}}{8\kappa_f} > 0$, then

$$g'(m) = \frac{\ln(1+z) - \frac{z}{1+z}\frac{n+m}{2m}}{\ln^2(1+z)}.$$

Since $\ln(1+z) > \frac{z}{1+z}$ for $z > 0$, we know that $g'(n) > 0$, therefore, $m^\star < n$.

Let $m = n^s$ where $0 < s < 1$, we would like to have $g'(n^s) < 0$, i,e.,

$$\frac{\ln(1+z)}{\frac{z}{1+z}} < \frac{1+n^{1-s}}{2}.$$

so that $m^\star \in (n^s, n)$.

Since $\kappa_f > n^{\frac{1}{2}}$, we have $z = \frac{\sqrt{m}}{8\kappa_f} < \frac{1}{8}$, on the other hand, we have

$$[\frac{\ln(1+z)}{\frac{z}{1+z}} < \frac{1+n^{1-s}}{2}]'_z > 0.$$

Therefore, it suffices to have

$$n^{1-s} > 18\ln\frac{9}{8} - 1 \coloneqq c_0 > 1.$$

As a result, we have $m^\star \in (\frac{n}{c_0}, n)$, and

$$C_1(m^\star, \varepsilon) = \mathcal{O}(\frac{n+m^\star}{\ln(1+\frac{\sqrt{m^\star}}{8\kappa_f})}\ln\frac{1}{\varepsilon}) = \mathcal{O}(\frac{n}{\frac{\sqrt{n}}{8\kappa_f}}\ln\frac{1}{\varepsilon}) = \mathcal{O}(\kappa_f\sqrt{n}\ln\frac{1}{\varepsilon}),$$

where in the second equality we have used $\kappa_f > n^{\frac{1}{2}}$.

For our iPreSVRG in Algorithm 3.1, we have

$$C_1'(m, \varepsilon) = \mathcal{O}(\frac{n+(1+pd)m}{\ln(1+\frac{1}{4}m\eta\sigma^M)}\ln\frac{1}{\varepsilon}),$$

thanks to Lemma 3.4.4, p can be chosen as

$$p = \mathcal{O}(\sqrt{\kappa(M)}\ln\left(\sqrt{\kappa_f^M}\kappa(M)\right),$$

furthermore, we can take $\eta = \frac{1}{2\sqrt{m}L_f}$ due to Theorem 3.4.2.

Under these settings, we have

$$C_1'(m, \varepsilon) = \mathcal{O}(\frac{n+(1+pd)m}{\ln(1+\frac{1}{8}\frac{\sqrt{m}}{\kappa_f^M})}\ln\frac{1}{\varepsilon}).$$

Let us take $m = m' = \lceil \frac{n}{1+pd} \rceil$.

If $n > 1 + pd$, or equivalently $\kappa_f < n^2 d^{-2}$, then

$$C_1'(m', \varepsilon) = \mathcal{O}(\frac{n}{\ln(1 + \frac{1}{8}\frac{\sqrt{n}}{\sqrt{pd}\kappa_f^M})} \ln \frac{1}{\varepsilon}).$$

Since $p = \mathcal{O}\left(\sqrt{\kappa(M)} \ln\left(\sqrt{\kappa_f^M}\kappa(M)\right)\right)$, we know that when $(\kappa_f^M)^2\sqrt{\kappa(M)}d < n$, or equivalently $\kappa_f < n^2 d^{-2}$, we have

$$\ln(1 + \frac{1}{8}\frac{\sqrt{n}}{\sqrt{pd}\kappa_f^M}) = \mathcal{O}(\ln n),$$

therefore

$$C_1'(m', \varepsilon) = \mathcal{O}(n \ln \frac{1}{\varepsilon}),$$

and

$$\frac{\min_{m\geq 1} C_1'(m, \varepsilon)}{\min_{m\geq 1} C_1(m, \varepsilon)} \leq \frac{C_1'(m', \varepsilon)}{C_1(m^\star, \varepsilon)} = \mathcal{O}(\frac{\sqrt{n}}{\kappa_f}).$$

If $n \leq 1 + pd$, or equivalently $\kappa_f > n^2 d^{-2}$, then $m = 1$ and

$$C_1'(m, \varepsilon) = \mathcal{O}(\frac{\sqrt{\kappa(M)}d}{\ln(1 + \frac{1}{8}\frac{1}{\kappa_f^M})} \ln \frac{1}{\varepsilon}),$$

therefore

$$\frac{\min_{m\geq 1} C_1'(m, \varepsilon)}{\min_{m\geq 1} C_1(m, \varepsilon)} \leq \frac{C_1'(1, \varepsilon)}{C_1(m^\star, \varepsilon)} = \mathcal{O}(\frac{\sqrt{\kappa(M)}d}{\kappa_f\sqrt{n}\ln(1 + \frac{1}{8}\frac{1}{\kappa_f^M})}).$$

Since $\kappa(M) \approx \kappa_f \gg \kappa_f^M$, this ratio becomes $\mathcal{O}(\frac{d}{\sqrt{n\kappa_f}})$ □

Proof of Theorem 3.4.6. The proof of Theorem 3.4.6 is similar and is omitted. □

Part III

Convergence Behaviors on Pathological Problems

In this part, we present the results of [137] and [199], where the behavior of DRS on pathological convex problems is analyzed. This part depends heavily on the monotone operator theory (c.f. [18]).

In Chapter 4, we work on conic programs. First, we view DRS iteration as a fixed-point iteration with some firmly nonexpansive operator T (see (4.6)). In the pathological settings, T does not have a fixed point, and DRS iterations will diverge. However, they diverge in a certain pattern 4.2.3, and this pattern is very helpful for identifying pathologies. Specifically, we can run three different but similar fixed-point iterations in parallel (see Algorithms 4.1, 4.2, and 4.3). Their convergence or divergence patterns inform us about what goes wrong in the original conic program, and how we may fix them. We summarize the theoretical results as a flowchart for identifying pathologies (see Figure 4.1), and numerical results on infeasible semidefinite programs (SDPs) in Section 4.3.

In Chapter 5, we turn to general convex problems. Just like DRS for conic programs, the divergence pattern of DRS can still inform us about certain pathologies such as strong infeasibility and improving directions (see Section 5.3.1). Furthermore, we show in Section 5.3.2 that, DRS essentially only requires strong duality to "work" even when the primal and/or dual solution does not exist, in the sense that the objective values of the iterates are asymptotically optimal. This result comes from a novel function value analysis. Finally, all these results are translated for ADMM in Section 5.5, which is known to be equivalent to DRS.

CHAPTER 4

DRS for Pathological Conic Programs

4.1 Introduction

Many convex optimization algorithms have strong theoretical guarantees and empirical performance, but they are often limited to non-pathological, feasible problems; under pathologies often the theory breaks down and the empirical performance degrades significantly. In fact, the behavior of convex optimization algorithms under pathologies has been studied much less, and many existing solvers often simply report "failure" without informing the users of what went wrong upon encountering infeasibility, unboundedness, or pathology. Pathological problem are numerically challenging, but they are not impossible to deal with. As infeasibility, unboundedness, and pathology can arise in practice (see, for example, [141, 140, 225, 229, 78]), designing a robust algorithm that behaves well in all cases is important to the completion of a robust solver.

In this chapter, we propose a method based on Douglas-Rachford splitting (DRS) that identifies infeasible, unbounded, and pathological conic programs. First-order methods such as DRS are simple and can quickly provide a solution with low or moderate accuracy. It is well known, for example by combining Theorem 1 of [193] and Proposition 4.4 of [82], that the iterates of DRS converge to a fixed point if there is one (a fixed point z^* of an operator T satisfies $z^* = Tz^*$), and when there is no fixed point, the iterates diverge unboundedly. However, the precise manner in which they diverge has been studied much less. Somewhat surprisingly, when iterates of DRS diverge, the divergent iterates still provide useful information, which we use to classify the conic

program. For example, a separating hyperplane can be found when the conic program is strongly infeasible, and an improving direction can be obtained when there is one. When the problem is infeasible or weakly feasible, we can get information of how to minimally modify the problem data to achieve strong feasibility.

Facial reduction is one approach to handle infeasible or pathological conic programs. Facial reduction reduces an infeasible or pathological problem into a reduced problem that is strongly feasible, strongly infeasible, or unbounded with an improving direction, which are the easier cases [38, 36, 169, 227]. This reduced problem can then be solved with, say, interior point methods [162]. However, facial reduction introduces a new set of computational issues. After completing the facial reduction step, which has its own the computational challenge and cost, the reduced problem must be solved. The reduced problem involves a cone expressed as an intersection of the original cone with an linear subspace, and in general such cones neither are self-dual nor have a simple formula for projection. This makes applying an interior point method or a first-order method difficult, and existing work on facial reduction do not provide an efficient way to address this issue.

Homogeneous self-dual embedding is a transformation that embeds a conic program and its dual into a single larger conic program. In conjunction with interior point methods, one can use the homogeneous self-dual embedding to identify and solve some pathologies [240, 70, 241, 143, 175].

In contrast, our proposed method directly addresses infeasibility, unboundedness, and pathology without transforming to a larger problem. Some cases are always identified, and some are identifiable under certain conditions. Being a first-order method, the proposed algorithm relies on simple subroutines; each iteration performs projections onto the cone and the affine space of the conic program and elementary operations such as vector addition. Consequently, the method is simple to implement and has a lower per-iteration cost than interior point methods.

4.1.1 Basic definitions

Cones. A set $K \subseteq \mathbb{R}^n$ is a cone if $K = \lambda K$ for any $\lambda > 0$. We write and define the dual cone of K as

$$K^* = \{u \in \mathbb{R}^n | \; u^T v \geq 0, \text{ for all } v \in K\}.$$

Throughout this chapter, we focus on nonempty closed convex cones that we can efficiently project onto. In particular, we do *not* require that the cone be self-dual. Example of such cones include:

- The positive orthant:

$$\mathbb{R}^k_+ = \{x \in \mathbb{R}^k \,|\, x_i \geq 0, \, i = 1, \ldots, n\}$$

- Second order cone:

$$Q^{k+1} = \left\{(x_1, \ldots, x_k, x_{k+1}) \in \mathbb{R}^k \times \mathbb{R}_+ \,|\, x_{k+1} \geq \sqrt{x_1^2 + \cdots + x_k^2}\right\}$$

- Rotated second order cone:

$$Q_r^{k+2} = \left\{(x_1, \ldots, x_k, x_{k+1}, x_{k+2}) \in \mathbb{R}^k \times \mathbb{R}^2_+ \,|\, 2x_{k+1}x_{k+2} \geq x_1^2 + \cdots + x_k^2\right\}.$$

- Positive semidefinite cone:

$$S^k_+ = \{M = M^T \in \mathbb{R}^{k \times k} | \; x^T M x \geq 0 \text{ for any } x \in \mathbb{R}^k\}$$

Conic programs. Consider the conic program

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & x \in K, \end{array} \tag{P}$$

where $x \in \mathbb{R}^n$ is the optimization variable, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$ are problem data, and $K \subseteq \mathbb{R}^n$ is a nonempty closed convex cone. We write $p^\star = \inf\{c^T x \,|\, Ax =$

$b,\ x \in K\}$ to denote the optimal value of (P). For simplicity, we assume $m \leq n$ and A is full rank.

The dual problem of (P) is

$$\begin{array}{ll} \text{maximize} & b^T y \\ \text{subject to} & A^T y + s = c \\ & s \in K^*, \end{array} \tag{D}$$

where $y \in \mathbb{R}^m$ and $s \in \mathbb{R}^n$ are the optimization variables. We write $d^\star = \sup\{b^T y \,|\, A^T y + s = c,\ s \in K^*\}$ to denote the optimal value of (D).

The optimization problem (P) is either feasible or infeasible; (P) is feasible if there is an $x \in K \cap \{x \,|\, Ax = b\}$ and infeasible if there is not. When (P) is feasible, it is strongly feasible if there is an $x \in \textbf{relint} K \cap \{x \,|\, Ax = b\}$ and weakly feasible if there is not, where **relint** denotes the relative interior. When (P) is infeasible, it is strongly infeasible if there is a non-zero distance between K and $\{x \,|\, Ax = b\}$, i.e., $d(K, \{x \,|\, Ax = b\}) > 0$, and weakly infeasible if $d(K, \{x \,|\, Ax = b\}) = 0$, where

$$d(C_1, C_2) = \inf \{\|x - y\| \,|\, x \in C_1,\ y \in C_2\},$$

and $\|\cdot\|$ denotes the Euclidean norm. Note that $d(C_1, C_2) = 0$ does not necessarily imply C_1 and C_2 intersect. When (P) is infeasible we say $p^\star = \infty$ and when feasible $p^\star \in \mathbb{R} \cup \{-\infty\}$. Likewise, when (D) is infeasible we say $d^\star = -\infty$ and when feasible $d^\star \in \mathbb{R} \cup \{\infty\}$.

As special cases, (P) is called a linear program when K is the positive orthant, a second-order cone program when K is the second-order cone, and a semidefinite program when K is the positive semidefinite cone.

4.1.2 Classification of conic programs

Every conic program of the form (P) falls under exactly one of the following 7 cases (some of the following examples are taken from [146, 148, 143, 145]). Discussions on

most of these cases exist in the literature. Some of these cases have a corresponding dual characterization, but we skip this discussion as it is not directly relevant to our method. We report the results of SDPT3 [220], SeDuMi [211], and MOSEK [156] using their default settings. In Section 4.2, we discuss how to identify most of these 7 cases.

Case (a). $p^\star$ is finite, both (P) and (D) have solutions, and $d^\star = p^\star$, which is the most common case. For example, the problem

$$\begin{array}{ll} \text{minimize} & x_3 \\ \text{subject to} & x_1 = 1 \\ & x_3 \geq \sqrt{x_1^2 + x_2^2} \end{array}$$

has the solution $x^\star = (1, 0, 1)$ and $p^\star = 1$. (The inequality constraint corresponds to $x \in Q^3$.) SDPT3, SeDuMi and MOSEK can solve this example.

The dual problem, after some simplification, is

$$\begin{array}{ll} \text{maximize} & y \\ \text{subject to} & 1 \geq y^2, \end{array}$$

which has the solution $y^\star = 1$ and $d^\star = 1$.

Case (b). $p^\star$ is finite, (P) has a solution, but (D) has no solution, $d^\star < p^\star$, or both. For example, the problem

$$\begin{array}{ll} \text{minimize} & x_2 \\ \text{subject to} & x_1 = x_3 = 1 \\ & x_3 \geq \sqrt{x_1^2 + x_2^2} \end{array}$$

has the solution $x^\star = (1, 0, 1)$ and optimal value $p^\star = 0$. (The inequality constraint corresponds to $x \in Q^3$.)

The dual problem, after some simplification, is

$$\text{maximize} \quad y_1 - \sqrt{1 + y_1^2}.$$

By taking $y_1 \to \infty$, we achieve the dual optimal value $d^\star = 0$, but no finite y_1 achieves it.

In this example, SDPT3 reports "Inaccurate/Solved" and -2.99305×10^{-5} as the optimal value; SeDuMi reports "Solved" and -1.54566×10^{-4} as the optimal value; MOSEK reports "Solved" and -2.71919×10^{-8} as the optimal value.

As another example, the problem

$$\begin{array}{ll} \text{minimize} & 2x_{12} \\ \text{subject to} & X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{12} & 0 & x_{23} \\ x_{13} & x_{23} & x_{12}+1 \end{bmatrix} \in S^3_+, \end{array}$$

has the solution

$$X^\star = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and optimal value $p^\star = 0$.

The dual problem, after some simplification, is

$$\begin{array}{ll} \text{maximize} & 2y_2 \\ \text{subject to} & \begin{bmatrix} 0 & y_2+1 & 0 \\ y_2+1 & -y_1 & 0 \\ 0 & 0 & -2y_2 \end{bmatrix} \in S^3_+, \end{array}$$

which has the solution $y^\star = (0, -1)$ and optimal value $d^\star = -2$.

In this example, SDPT3 reports "Solved" and -2 as the optimal value; SeDuMi reports "Solved" and -0.602351 as the optimal value; MOSEK reports "Failed" and does not report an optimal value.

Note that case (b) can happen only when (P) is weakly feasible, by standard convex duality [191].

Case (c). (P) is feasible, $p^\star$ is finite, but there is no solution. For example, the problem

$$\begin{array}{ll} \text{minimize} & x_3 \\ \text{subject to} & x_1 = \sqrt{2} \\ & 2x_2x_3 \geq x_1^2 \\ & x_2, x_3 \geq 0 \end{array}$$

has an optimal value $p^\star = 0$ but has no solution since any feasible x satisfies $x_3 > 0$. (The inequality constraints correspond to $x \in Q_r^3$.)

In this example, SDPT3 reports "Inaccurate/Solved" and 7.9509×10^{-5} as the optimal value; SeDuMi reports "Solved" and 8.75436×10^{-5} as the optimal value; MOSEK reports "Solved" and 4.07385×10^{-8} as the optimal value.

Case (d). (P) is feasible, $p^\star = -\infty$, and there is an improving direction, i.e., there is a $u \in \mathcal{N}(A) \cap K$ satisfying $c^T u < 0$. For example, the problem

$$\begin{array}{ll} \text{minimize} & x_1 \\ \text{subject to} & x_2 = 0 \\ & x_3 \geq \sqrt{x_1^2 + x_2^2} \end{array}$$

has an improving direction $u = (-1, 0, 1)$. If x is any feasible point, $x + tu$ is feasible for $t \geq 0$, and the objective value goes to $-\infty$ as $t \to \infty$. (The inequality constraint corresponds to $x \in Q^3$.)

In this example, SDPT3 reports "Failed" and does not report an optimal value; SeDuMi reports "Unbounded" and $-\infty$ as the optimal value; MOSEK reports "Unbounded" and $-\infty$ as the optimal value.

Case (e). (P) is feasible, $p^\star = -\infty$, but there is no improving direction, i.e., there is no $u \in \mathcal{N}(A) \cap K$ satisfying $c^T u < 0$. For example, consider the problem

$$\begin{array}{ll} \text{minimize} & x_1 \\ \text{subject to} & x_2 = 1 \\ & 2x_2 x_3 \geq x_1^2 \\ & x_2, x_3 \geq 0. \end{array}$$

(The inequality constraints correspond to $x \in Q_r^3$.) Any improving direction $u = (u_1, u_2, u_3)$ would satisfy $u_2 = 0$, and this in turn, with the cone constraint, implies $u_1 = 0$ and $c^T u = 0$. However, even though there is no improving direction, we can eliminate the variables x_1 and x_2 to verify that

$$p^\star = \inf\{-\sqrt{2x_3} \,|\, x_3 \geq 0\} = -\infty.$$

In this example, SDPT3 reports "Failed" and does not report an optimal value; SeDuMi reports "Inaccurate/Solved" and -175514 as the optimal value; MOSEK reports "Inaccurate/Unbounded" and $-\infty$ as the optimal value.

Case (f). Strongly infeasible, where $p^\star = \infty$ and $d(K, \{x \,|\, Ax = b\}) > 0$. For example, the problem

$$\begin{array}{ll} \text{minimize} & 0 \\ \text{subject to} & x_3 = -1 \\ & x_3 \geq \sqrt{x_1^2 + x_2^2} \end{array}$$

satisfies $d(K, \{x \,|\, Ax = b\}) = 1$. (The inequality constraint corresponds to $x \in Q^3$.)

In this example, SDPT3 reports "Failed" and does not report an optimal value; SeDuMi reports "Infeasible" and ∞ as the optimal value; MOSEK reports "Infeasible" and ∞ as the optimal value.

Case (g). Weakly infeasible, where $p^\star = \infty$ but $d(K, \{x \mid Ax = b\}) = 0$. For example, the problem

$$\begin{array}{ll} \text{minimize} & 0 \\ \text{subject to} & \begin{bmatrix} 0, 1, 1 \\ 1, 0, 0 \end{bmatrix} x = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ & x_3 \geq \sqrt{x_1^2 + x_2^2} \end{array}$$

satisfies $d(K, \{x \mid Ax = b\}) = 0$, since

$$d(K, \{x \mid Ax = b\}) \leq \|(1, -y, y) - (1, -y, \sqrt{y^2 + 1})\| \to 0$$

as $y \to \infty$. (The inequality constraint corresponds to $x \in Q^3$.)

In this example, SDPT3 reports "Infeasible" and ∞ as the optimal value; SeDuMi reports "Solved" and 0 as the optimal value; MOSEK reports "Failed" and does not report an optimal value.

Remark. In the case of linear programming, i.e., when K in (P) is the positive orthant, there are only three possible cases: (a), (d), and (f).

4.1.3 Classification method overview

At a high level, our proposed method for classifying the 7 cases is quite simple. Given an operator T and a starting point z^0, we call $z^{k+1} = T(z^k)$ the *fixed-point iteration* of T. Our proposed method runs three similar but distinct fixed-point iterations with the operators

$$\begin{aligned} T_1(z) &= \tilde{T}(z) + x_0 - \gamma Dc \\ T_2(z) &= \tilde{T}(z) + x_0 \\ T_3(z) &= \tilde{T}(z) - \gamma Dc, \end{aligned} \qquad \text{(Operators)}$$

where $\tilde{T}(z) = (1/2)(I + R_{\mathcal{N}(A)} R_K)(z)$, $D = I - A^T(AA^T)^{-1}A$, $x_0 = A^T(AA^T)^{-1}b$, and $\gamma > 0$. We explain the notation in more detail in Section 4.2.

We can view T_1 as the DRS operator of (P), T_2 as the DRS operator with c set to $\mathbf{0}$ in (P), and T_3 as the DRS operator with b set to $\mathbf{0}$ in (P). We use the information provided by the iterates of these fixed-point iterations to solve (P) and classify the cases. As outlined in Section 4.2.8, this is based on the theory of Section 4.2 and the flowchart shown in Figure 4.1.

4.1.4 Previous work

Previously, Bauschke, Combettes, Hare, Luke, and Moursi have analyzed Douglas-Rachford splitting in other pathological problems such as: feasibility problems between 2 affine sets [25], feasibility problems between 2 convex sets [19, 26], and general setups [13, 21, 23, 157]. Our work builds on these past results.

4.2 Obtaining certificates from Douglas-Rachford Splitting

The primal problem (P) is equivalent to

$$\text{minimize} \quad f(x) + g(x), \tag{4.1}$$

where

$$\begin{aligned} f(x) &= c^T x + \delta_{\{x \,|\, Ax=b\}}(x) \\ g(x) &= \delta_K(x), \end{aligned} \tag{4.2}$$

and $\delta_C(x)$ is the indicator function of a set C defined as

$$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C. \end{cases}$$

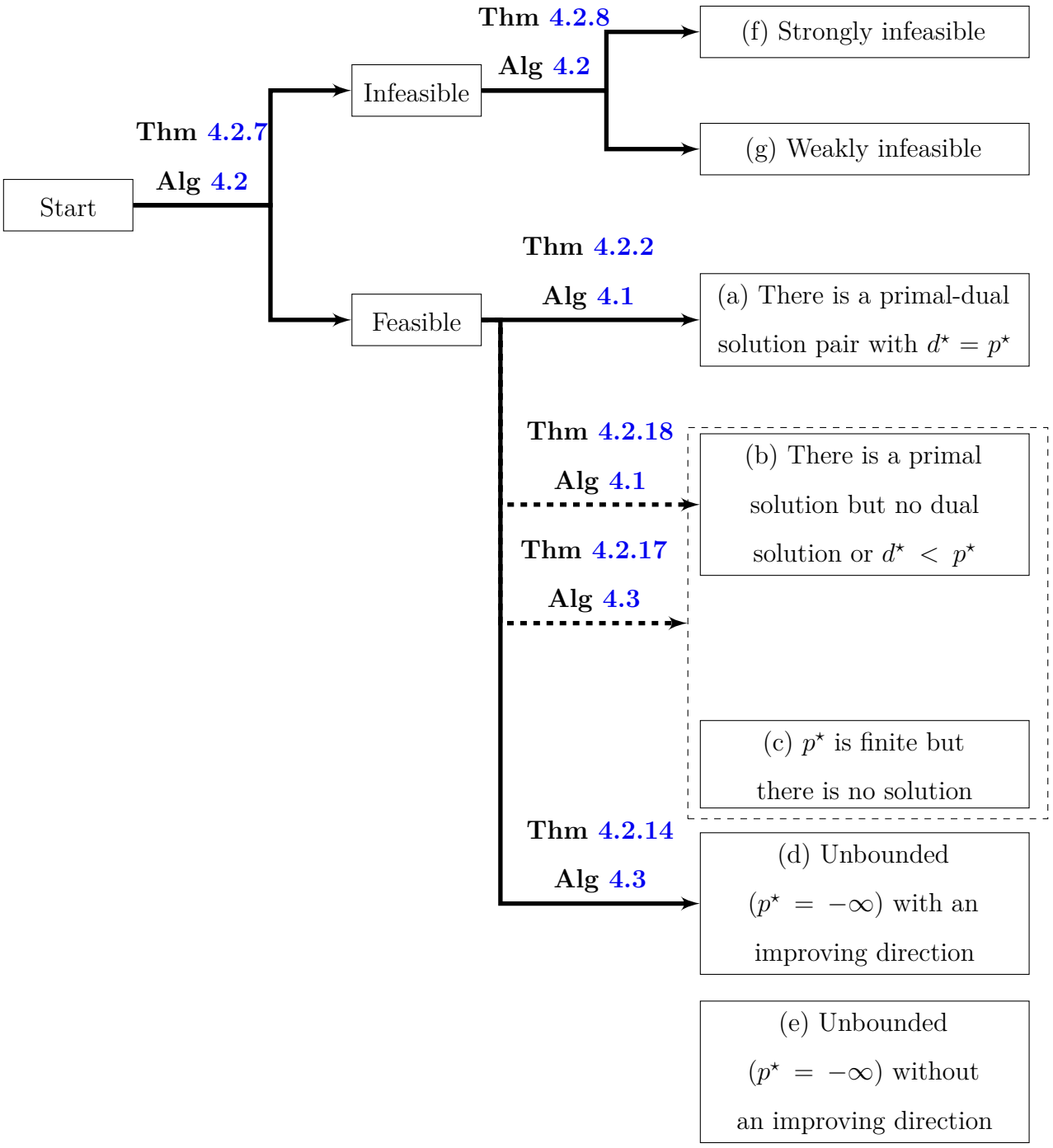

Figure 4.1: The flowchart for identifying cases (a)–(g). A solid arrow means the cases are always identifiable, a dashed arrow means the cases sometimes identifiable.

Douglas-Rachford splitting (DRS) [133] applied to (3.32) is

$$\begin{aligned} x^{k+1/2} &= \text{Prox}_{\gamma g}(z^k) \\ x^{k+1} &= \text{Prox}_{\gamma f}(2x^{k+1/2} - z^k) \\ z^{k+1} &= z^k + x^{k+1} - x^{k+1/2}, \end{aligned} \tag{4.3}$$

which updates z^k to z^{k+1} for $k = 0, 1, \ldots$. Given $\gamma > 0$ and function h,

$$\text{Prox}_{\gamma h}(x) = \underset{z \in \mathbb{R}^n}{\arg\min} \left\{ h(z) + (1/2\gamma)\|z - x\|^2 \right\}$$

denotes the proximal operator with respect to γh.

Given a nonempty closed convex set $C \subseteq \mathbb{R}^n$, define the projection with respect to C as

$$P_C(x) = \underset{y \in C}{\arg\min} \|y - x\|^2$$

and the reflection with respect to C as

$$R_C(x) = 2P_C(x) - x.$$

Write I to denote both the $n \times n$ identity matrix and the identity map from $\mathbb{R}^n \to \mathbb{R}^n$. Write $\mathbf{0}$ to denote the origin point in $\mathbb{R}^n$. Define

$$\begin{aligned} D &= I - A^T(AA^T)^{-1}A \\ x_0 &= A^T(AA^T)^{-1}b = P_{\{x \,|\, Ax = b\}}(\mathbf{0}). \end{aligned} \tag{4.4}$$

Write $\mathcal{N}(A)$ for the null space of A and $\mathcal{R}(A^T)$ for the range of A^T. Then

$$\begin{aligned} P_{\{x \,|\, Ax = b\}}(x) &= Dx + x_0, \\ P_{\mathcal{N}(A)}(x) &= Dx. \end{aligned}$$

Finally, define

$$\tilde{T}(z) = \frac{1}{2}(I + R_{\mathcal{N}(A)}R_K)(z).$$

Now we can rewrite the DRS iteration (4.3) as

$$\begin{aligned} x^{k+1/2} &= P_K(z^k) \\ x^{k+1} &= D(2x^{k+1/2} - z^k) + x_0 - \gamma Dc \\ z^{k+1} &= z^k + x^{k+1} - x^{k+1/2}. \end{aligned} \tag{4.5}$$

Equivalently and more compactly, we can write

$$z^{k+1} = \tilde{T}(z^k) + x_0 - \gamma Dc, \tag{4.6}$$

which is also $z^{k+1} = T_1(z^k)$ with T_1 definied in (Operators).

Remark. Instead of (3.33), we could have considered the more general form

$$\begin{aligned} f(x) &= (1-\alpha)c^T x + \delta_{\{x \,|\, Ax=b\}}(x), \\ g(x) &= \alpha c^T x + \delta_K(x) \end{aligned}$$

with $\alpha \in \mathbb{R}$. By simplifying the resulting DRS iteration, one can verify that the iterates are equivalent to the $\alpha = 0$ case. Since the choice of α does not affect the DRS iteration at all, we will only work with the case $\alpha = 0$.

4.2.1 Convergence of DRS

A point $x^\star \in \mathbb{R}^n$ is a solution of (3.32) if and only if

$$\mathbf{0} \in \partial(f+g)(x^\star).$$

DRS, however, converges if and only if there is a point $x^\star$ such that

$$\mathbf{0} \in \partial f(x^\star) + \partial g(x^\star).$$

In general,

$$\partial f(x) + \partial g(x) \subseteq \partial(f+g)(x)$$

for all $x \in \mathbb{R}^n$, but the two are not necessarily equal.

We summarize the convergence of DRS in the theorem below. Its main part is a direct result of Theorem 1 of [193] and Propositions 4.4 and 4.8 of [82]. The convergence of $x^{k+1/2}$ and x^{k+1} is due to [214]. Therefore, we do not prove it.

Theorem 4.2.1. *Consider the iteration (4.6) with any starting point z^0. If there is an x such that*

$$\mathbf{0} \in \partial f(x) + \partial g(x),$$

then z^k converges to a limit $z^\star$, $x^{k+1/2} \to x^\star = \mathrm{Prox}_{\gamma g}(z^\star)$, $x^{k+1} \to x^\star = \mathrm{Prox}_{\gamma g}(z^\star)$, and

$$\mathbf{0} \in \partial f(x^\star) + \partial g(x^\star).$$

If there is no x such that

$$\mathbf{0} \in \partial f(x) + \partial g(x),$$

then z^k diverges in that $\|z^k\| \to \infty$.

DRS can fail to find a solution to (P) even when one exists. Slater's constraint qualification is a sufficient condition that prevents such pathologies: if (P) is strongly feasible, then

$$\mathbf{0} \in \partial f(x^\star) + \partial g(x^\star)$$

for all solutions $x^\star$ [190, Theorem 23.8]. This fact and Theorem 4.2.1 tell us that under Slater's constraint qualifications DRS finds a solution of (P) if one exists.

The following theorem, however, provides a stronger, necessary and sufficient characterization of when the DRS iteration converges.

Theorem 4.2.2 ([191])**.** *There is an $x^\star$ such that*

$$\mathbf{0} \in \partial f(x^\star) + \partial g(x^\star)$$

if and only if $x^\star$ is a solution to (P), (D) has a solution, and $d^\star = p^\star$.

Based on Theorem 4.2.1 and 4.2.2 we can determine whether we have case (a) with the iteration (4.6)

with any starting point z^0 and $\gamma > 0$.

- If $\lim_{k\to\infty} \|z^k\| < \infty$, we have case (a), and vice versa.
- If $\lim_{k\to\infty} \|z^k\| = \infty$, we do not have case (a), and vice versa.

With a finite number of iterations, we test $\|z^k\| \geq M$ for some large $M > 0$. However, distinguishing the two cases can be numerically difficult as the rate of $\|z^k\| \to \infty$ can be very slow.

4.2.2 Fixed-point iterations without fixed points

We say an operator $T : \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive if

$$\|T(x) - T(y)\|^2 \leq \|x - y\|^2$$

for all $x, y \in \mathbb{R}^n$. We say T is firmly nonexpansive (FNE) if

$$\|T(x) - T(y)\|^2 \leq \|x - y\|^2 - \|(I - T)(x) - (I - T)(y)\|^2$$

for all $x, y \in \mathbb{R}^n$. (FNE operators are nonexpansive.) In particular, all three operators defined in (Operators) are FNE, as they are DRS operators [15]. It is well known [66] that if a FNE operator T has a fixed point, its fixed-point iteration $z^{k+1} = T(z^k)$ converges to one with rate

$$\|z^{k+1} - z^k\| = o(1/\sqrt{k}).$$

Now consider the case where a FNE operator T has no fixed point, which has been studied to a lesser extent. In this case, the fixed-point iteration $z^{k+1} = T(z^k)$ diverges in that $\|z^k\| \to \infty$ [193, Theorem 1]. Precisely in what manner z^k diverges is characterized by the *infimal displacement vector* [172]. Given a FNE operator T, we call

$$v = P_{\overline{\mathbf{ran}\,(I-T)}}(\mathbf{0})$$

the *infimal displacement vector* of T. To clarify, $\overline{\mathbf{ran}\,(I-T)}$ denotes the closure of the set

$$\mathbf{ran}\,(I-T) = \{x - T(x) \,|\, x \in \mathbb{R}^n\}.$$

Because T is nonexpansive, the closed set $\overline{\mathbf{ran}\,(I-T)}$ is convex [172], so v is uniquely defined. We can interpret the infimal displacement vector v as the asymptotic output of $I - T$ corresponding to the best effort to find a fixed point.

Lemma 4.2.3 (Corollary 2.3 of [11])**.** *Let T be FNE, and consider its fixed-point iteration $z^{k+1} = T(z^k)$ with any starting point z^0. Then*

$$z^k - z^{k+1} \to v = P_{\overline{\mathbf{ran}\,(I-T)}}(\mathbf{0}).$$

In [11], Lemma 4.2.3 is proved in generality for nonexpansive operators, but we provide a simpler proof in our setting in Theorem 4.2.4.

When T has a fixed point $v = \mathbf{0}$, but $v = \mathbf{0}$ is possible even when T has no fixed point. In the following sections, we use Lemma 4.2.3 to determine the status of a conic program, but, in general, $z^k - z^{k+1} \to v$ has no rate. However, we only need to determine whether $\lim_{k\to\infty}(z^{k+1} - z^k) = \mathbf{0}$ or $\lim_{k\to\infty}(z^{k+1} - z^k) \neq \mathbf{0}$, and we do so by checking whether $\|z^{k+1} - z^k\| \geq \varepsilon$ for some tolerance $\varepsilon > 0$. For this purpose, the following rate of approximate convergence is good enough.

Theorem 4.2.4. *Let T be FNE, and consider its fixed point iteration*

$$z^{k+1} = T(z^k),$$

with any starting point z^0, then

$$z^k - z^{k+1} \to v.$$

And for any $\varepsilon > 0$, there is an $M_\varepsilon > 0$ (which depends on T, z^0, and ε) such that

$$\|v\| \leq \min_{0\leq j\leq k} \|z^j - z^{j+1}\| \leq \|v\| + \frac{M_\varepsilon}{\sqrt{k+1}} + \frac{\varepsilon}{2}.$$

Proof of Theorem 4.2.4. For simplicity, we prove the result for $0 < \varepsilon \leq 1$. The result for $\varepsilon = 1$ applies to the $\varepsilon > 1$ case.

Given any x_ε, we use the triangle inequality to get

$$\|z^k - z^{k+1} - v\| = \|T^k(z^0) - T^{k+1}(z^0) - v\| \tag{4.7}$$

$$\leq \|(T^k(z^0) - T^{k+1}(z^0)) - (T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon))\| + \|T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon) - v\|. \tag{4.8}$$

To bound the second term, pick an x_ε such that

$$\|x_\varepsilon - T(x_\varepsilon) - v\| \leq \frac{\varepsilon^2}{4(2\|v\| + 1)},$$

which we can do since $v = P_{\overline{\mathbf{ran}\,(I-T)}}(\mathbf{0}) \in \overline{\mathbf{ran}\,(I-T)}$. Since T is nonexpansive, we have

$$\|T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon)\| - \|v\| \leq \|x_\varepsilon - T(x_\varepsilon)\| - \|v\| \leq \|x_\varepsilon - T(x_\varepsilon) - v\|.$$

Since $v = \arg\min_{\overline{\mathbf{ran}\,(I-T)}} \|x\|$, we have $\|T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon)\| - \|v\| \geq 0$. Putting this together we get

$$0 \leq \|T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon)\| - \|v\| \leq \frac{\epsilon^2}{4(2\|v\| + 1)}.$$

Since $v = P_{\overline{\mathbf{ran}\,(I-T)}}(\mathbf{0})$,

$$\|v\|^2 \leq y^T v$$

for any $y \in \overline{\mathbf{ran}\,(I-T)}$. Putting these together we get

$$\begin{aligned}
\|T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon) - v\|^2 &= \|T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon)\|^2 + \|v\|^2 - 2(T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon))^T v \\
&\leq \|T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon)\|^2 + \|v\|^2 - 2\|v\|^2 \\
&= (\|T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon)\| + \|v\|)(\|T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon)\| - \|v\|) \\
&\leq (2\|v\| + \frac{\varepsilon^2}{4(2\|v\| + 1)})\frac{\varepsilon^2}{4(2\|v\| + 1)} \\
&\leq (2\|v\| + 1)\frac{\varepsilon^2}{4(2\|v\| + 1)} = \frac{\varepsilon^2}{4}
\end{aligned} \tag{4.9}$$

for $0 < \varepsilon \leq 1$.

Now let us bound the first term $\|(T^k(z^0) - T^{k+1}(z^0)) - (T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon))\|$ on the righthand side of (4.8). Since T is FNE, we have

$$\|(T^k(z^0)-T^{k+1}(z^0))-(T^k(x_\varepsilon)-T^{k+1}(x_\varepsilon))\|^2 = \|T^k(z^0)-T^k(x_\varepsilon)\|^2-\|T^{k+1}(z^0)-T^{k+1}(x_\varepsilon)\|^2.$$

Summing this inequality we have

$$\sum_{j=0}^{k} \|(T^k(z^0) - T^{k+1}(z^0)) - (T^k(x_\varepsilon) - T^{k+1}(x_\varepsilon))\|^2 \leq \|z^0 - x_\varepsilon\|^2. \tag{4.10}$$

(4.8), (4.9), and (4.10) imply that

$$z^k - z^{k+1} \to v.$$

Furthermore,

$$\min_{0\leq j\leq k} \|z^j - z^{j+1} - v\| \leq \frac{M_\varepsilon}{\sqrt{k+1}} + \frac{\varepsilon}{2},$$

where $M_\varepsilon = \|z^0 - x_\varepsilon\|$. As a result,

$$\|v\| \leq \min_{0\leq j\leq k} \|z^j - z^{j+1}\| \leq \|v\| + \frac{M_\varepsilon}{\sqrt{k+1}} + \frac{\varepsilon}{2}.$$

□

4.2.3 Feasibility and infeasibility

We now return to conic programs. Consider the operator T_2 defined by $T_2(z) = \tilde{T}(z)+x_0$. As mentioned, we can view T_2 as the DRS operator with c set to $\mathbf{0}$ in (P).

The infimal displacement vector of T_2 has a nice geometric interpretation: it is the best approximation displacement between the sets K and $\{x \,|\, Ax = b\}$, and $\|v\| = d(K, \{x \,|\, Ax = b\})$. Define the set

$$K - \{x \,|\, Ax = b\} = \{y - x \,|\, y \in K,\, Ax = b\}.$$

Theorem 4.2.5 (Theorem 3.4 of [19], Proposition 11.22 of [157])**.** *The operator T_2 has the infimal displacement vector $v = P_{\overline{K-\{x \,|\, Ax=b\}}}(\mathbf{0})$.*

We can further understand v in terms of the projection $P_{\overline{P_{\mathcal{R}(A^T)}(K)}}$. Note that $P_{\mathcal{R}(A^T)}(K)$ is a cone because K is. $P_{\mathcal{R}(A^T)}(K)$ is not always closed, but its closure $\overline{P_{\mathcal{R}(A^T)}(K)}$ is. We prove the following result at the end of this subsection.

Lemma 4.2.6 (Interpretation of v)**.** *The infimal displacement vector v of T_2 satisfies*

$$v = P_{\overline{K-\{x \,|\, Ax=b\}}}(\mathbf{0}) = P_{\overline{P_{\mathcal{R}(A^T)}(K)-x_0}}(\mathbf{0}) = P_{\overline{P_{\mathcal{R}(A^T)}(K)}}(x_0) - x_0,$$

where x_0 is given in (4.4) and K is any nonempty set.

Combining the discussion of Section 4.2.2 with Theorem 4.2.5 gives us Theorems 4.2.7 and 4.2.8.

Theorem 4.2.7 (Certificate of feasibility)**.** *Consider the iteration $z^{k+1} = T_2(z^k)$ with any starting point $z^0 \in \mathbb{R}^n$, then*

1. *(P) is feasible if and only if z^k converges, and in this case $x^{k+1/2}$ converges to a feasible point of (P).*
2. *(P) is infeasible if and only if z^k diverges in that $\|z^k\| \to \infty$.*

Theorem 4.2.8 (Certificate of strong infeasibility)**.** *Consider the iteration $z^{k+1} = T_2(z^k)$ with any starting point z^0. We have $z^k - z^{k+1} \to v$ and*

1. *(P) is strongly infeasible if and only if $v \neq \mathbf{0}$.*
2. *(P) is weakly infeasible or feasible if and only if $v = \mathbf{0}$.*

When (P) is strongly infeasible, we can obtain a separating hyperplane from v. We prove the following result at the end of this subsection.

Theorem 4.2.9 (Separating hyperplane)**.** *Consider the iteration $z^{k+1} = T_2(z^k)$ with any starting point z^0. When (P) is strongly infeasible, $z^k - z^{k+1} \to v \neq \mathbf{0}$, and the hyperplane*

$$\{x \,|\, h^T x = \beta\},$$

where $h = -v \in K^* \cap \mathcal{R}(A^T)$ *and* $\beta = -(v^T x_0)/2 > 0$*, strictly separates* K *and* $\{x \mid Ax = b\}$*. More precisely, for any* $y_1 \in K$ *and* $y_2 \in \{x \mid Ax = b\}$ *we have*

$$h^T y_1 < \beta < h^T y_2.$$

Based on Theorems 4.2.7, 4.2.8, and 4.2.9, we can determine feasibility, weak infeasiblity, and strong infeasibility and obtain a strictly separating hyperplane if one exists with the iteration $z^{k+1} = T_2(z^k)$ with any starting point z^0.

- $\lim_{k\to\infty} \|z^k\| < \infty$ if and only if (P) is feasible.
- $\lim_{k\to\infty} \|z^k - z^{k+1}\| > 0$ if and only if (P) is strongly infeasible, and Theorem 4.2.9 provides a strictly separating hyperplane.
- $\lim_{k\to\infty} \|z^k\| = \infty$ and $\lim_{k\to\infty} \|z^k - z^{k+1}\| = 0$ if and only if (P) is weakly infeasible.

With a finite number of iterations, we distinguish the three cases by testing $\|z^{k+1} - z^k\| \leq \varepsilon$ and $\|z^k\| \geq M$ for some small $\varepsilon > 0$ and large $M > 0$. By Theorem 4.2.4, we can distinguish strong infeasibility from weak infeasibility or feasibility at a rate of $O(1/\sqrt{k})$. However, distinguishing feasibility from weak infeasibility can be numerically difficult as the rate of $\|z^k\| \to \infty$ can be very slow when (P) is weakly infeasible.

Proof of Lemma 4.2.6. Remember that by definition (4.4), we have $x_0 \in \mathcal{R}(A^T)$ and

$$\{x \mid Ax = b\} = x_0 + \mathcal{N}(A) = x_0 - \mathcal{N}(A).$$

Also note that for any $y \in \mathbb{R}^n$, we have

$$y + \mathcal{N}(A) = P_{\mathcal{R}(A^T)}(y) + \mathcal{N}(A).$$

So

$$K - \{x \mid Ax = b\} = K + \mathcal{N}(A) - x_0 = P_{\mathcal{R}(A^T)}(K) - x_0 + \mathcal{N}(A),$$

and

$$\overline{K - \{x \mid Ax = b\}} = \overline{P_{\mathcal{R}(A^T)}(K) + \mathcal{N}(A)} - x_0 = \overline{P_{\mathcal{R}(A^T)}(K)} - x_0 + \mathcal{N}(A). \tag{4.11}$$

Since $x_0 \in \mathcal{R}(A^T)$, we have $\overline{P_{\mathcal{R}(A^T)}(K)} - x_0 \subseteq \mathcal{R}(A^T)$, and, in particular, $\overline{P_{\mathcal{R}(A^T)}(K)} - x_0$ is orthogonal to the subspace $\mathcal{N}(A)$. Recall

$$v = P_{\overline{P_{\mathcal{R}(A^T)}(K)} - x_0 + \mathcal{N}(A)}(\mathbf{0}).$$

So $v \in \overline{P_{\mathcal{R}(A^T)}(K)} - x_0 \subseteq \mathcal{R}(A^T)$ and

$$v = P_{\overline{P_{\mathcal{R}(A^T)}(K)} - x_0}(\mathbf{0}).$$

Finally,

$$v = \underset{x \in \overline{P_{\mathcal{R}(A^T)}(K)} - x_0}{\arg\min} \left\{ \|x\|_2^2 \right\} = \underset{y \in \overline{P_{\mathcal{R}(A^T)}(K)}}{\arg\min} \left\{ \|y - x_0\|_2^2 \right\} - x_0 = P_{\overline{P_{\mathcal{R}(A^T)}(K)}}(x_0) - x_0$$

□

Proof of Theorem 4.2.9. Note that

$$v = P_{\overline{K - \{x \mid Ax = b\}}}(\mathbf{0}) = P_{\overline{K + \mathcal{N}(A) - x_0}}(\mathbf{0}) = P_{\overline{K + \mathcal{N}(A)}}(x_0) - x_0$$

Using $I = P_{K^* \cap \mathcal{R}(A^T)} + P_{-(K^* \cap \mathcal{R}(A^T))^*}$ and $(K^* \cap \mathcal{R}(A^T))^* = \overline{K + \mathcal{N}(A)}$ [15], we have

$$v = P_{\overline{K + \mathcal{N}(A)}}(x_0) - x_0 = -P_{-(K^* \cap \mathcal{R}(A^T))}(x_0) = P_{K^* \cap \mathcal{R}(A^T)}(-x_0).$$

Since the projection operator is FNE, we have

$$-v^T x_0 = (v - \mathbf{0})^T(-x_0 - \mathbf{0}) \geq \|P_{K^* \cap \mathcal{R}(A^T)}(-x_0)\|^2 = \|v\|^2 > 0$$

and therefore $v^T x_0 < 0, \beta = -v^T x_0/2 > 0$.

So for any $y_1 \in K$ and $y_2 \in \{x \mid Ax = b\}$, we have

$$h^T y_1 = -v^T y_1 \leq 0 < -(v^T x_0)/2 = \beta < -v^T x_0 = h^T y_2,$$

where we have used $h = -v = -P_{K^* \cap \mathcal{R}(A^T)}(-x_0) \in -K^*$ in the first inequality. □

4.2.4 Modifying affine constraints to achieve strong feasibility

Loosely speaking, strongly feasible problems are the good cases that are easier to solve, compared to weakly feasible or infeasible problems. Given a problem that is not strongly feasible, how to minimally modify the problem to achieve strong feasibility is often useful to know.

The limit $z^k - z^{k+1} \to v$ informs us of how to do this. When $d(K, \{x \mid Ax = b\}) = \|v\| > 0$, the constraint $K \cap \{x \mid A(x-y) = b\}$ is infeasible for any y such that $\|y\| < \|v\|$. In general, the constraint $K \cap \{x \mid A(x-v) = b\}$ can be feasible or weakly infeasible, but is not strongly feasible. The constraint $K \cap \{x \mid A(x-v-d) = b\}$ is strongly feasible for an arbitrarily small $d \in \mathbf{relint} K$. In other words, $K \cap \{x \mid A(x-v-d) = b\}$ achieves strong feasibility with the minimal modification (measured by the Euclidean norm $\|\cdot\|$) to the original constraint $K \cap \{x \mid Ax = b\}$.

Theorem 4.2.10 (Achieving strong feasibility)**.** *Let $v = P_{\overline{K - \{x \mid Ax = b\}}}(\mathbf{0})$, and let d be any vector satisfying $d \in \mathbf{relint} K$. Then the constraint $K \cap \{x \mid A(x-v-d) = b\}$ is strongly feasible, i.e., there is an x such that $x \in \mathbf{relint} K \cap \{x \mid A(x-v-d) = b\}$.*

Proof of Theorem 4.2.10. By Lemma 4.2.6 we have

$$v + x_0 \in \overline{P_{\mathcal{R}(A^T)}(K)}. \tag{4.12}$$

Because $P_{\mathcal{R}(A^T)}$ is a linear transformation, by Lemma 4.2.11 below

$$P_{\mathcal{R}(A^T)}(\mathbf{relint} K) = \mathbf{relint} P_{\mathcal{R}(A^T)}(K).$$

Since $d \in \mathbf{relint} K$,

$$P_{\mathcal{R}(A^T)}(d) \in P_{\mathcal{R}(A^T)}(\mathbf{relint} K) = \mathbf{relint} P_{\mathcal{R}(A^T)}(K). \tag{4.13}$$

Applying Lemma 4.2.12 to (4.12) and (4.13), we have

$$v + x_0 + P_{\mathcal{R}(A^T)}(d) \in \mathbf{relint} P_{\mathcal{R}(A^T)}(K) = P_{\mathcal{R}(A^T)}(\mathbf{relint} K).$$

Finally we have

$$\mathbf{0} \in P_{\mathcal{R}(A^T)}(\mathbf{relint}K) - x_0 - v - d + \mathcal{N}(A) = \mathbf{relint}K - \{x \mid A(x - v - d) = b\}.$$

□

Lemma 4.2.11 (Theorem 6.6 of [190]). *If $A(\cdot)$ is a linear transformation and C is a convex set, then $A(\mathbf{relint}C) = \mathbf{relint}A(C)$.*

Lemma 4.2.12 (Theorem 6.1 [190]). *Let K be a convex cone. If $x \in K$ and $y \in \mathbf{relint}K$, then $x + y \in \mathbf{relint}K$.*

4.2.5 Improving direction

(P) has an improving direction if and only if the dual problem (D) is strongly infeasible:

$$0 < d(0, K^\star + \mathcal{R}(A^T) - c) = d(\{(y, s) \mid A^T y + s = c\}, \{(y, s) \mid s \in K^*\}).$$

Theorem 4.2.13 (Certificate of improving direction). *Exactly one of the following is true:*

1. *(P) has an improving direction, (D) is strongly infeasible, and $P_{\mathcal{N}(A)\cap K}(-c) \neq \mathbf{0}$ is an improving direction.*
2. *(P) has no improving direction, (D) is feasible or weakly infeasible, and $P_{\mathcal{N}(A)\cap K}(-c) = \mathbf{0}$.*

Furthermore,

$$P_{\mathcal{N}(A)\cap K}(-c) = P_{\overline{K^*+\mathcal{R}(A^T)-c}}(\mathbf{0}).$$

Theorem 4.2.14. *Consider the iteration $z^{k+1} = T_3(z^k) = \tilde{T}(z^k) - \gamma Dc$ with any starting point z^0 and $\gamma > 0$. If (P) has an improving direction, then*

$$d = \lim_{k\to\infty} z^{k+1} - z^k = P_{\overline{K^*+\mathcal{R}(A^T)-c}}(\mathbf{0}) \neq \mathbf{0}$$

gives one. If (P) *has no improving direction, then*

$$\lim_{k\to\infty} z^{k+1} - z^k = \mathbf{0}.$$

Based on Theorem 4.2.13 and 4.2.14 we can determine whether there is an improving direction and find one if one exists with the iteration $z^{k+1} = \tilde{T}(z^k) - \gamma Dc$ with any starting point z^0 and $\gamma > 0$.

- $\lim_{k\to\infty} z^{k+1} - z^k = \mathbf{0}$ if and only if there is no improving direction.
- $\lim_{k\to\infty} z^{k+1} - z^k = d \neq \mathbf{0}$ if and only if d is an improving direction.

With a finite number of iterations, we test $\|z^{k+1} - z^k\| \leq \varepsilon$ for some small $\varepsilon > 0$. By Theorem 4.2.4, we can distinguish whether there is an improving direction or not at a rate of $O(1/\sqrt{k})$.

We need the following theorem for Section 4.2.7, it is proved similarly to 4.2.7 below.

Theorem 4.2.15. *Consider the iteration*

$$z^{k+1} = \tilde{T}(z^k) - \gamma Dc$$

with any starting point z^0 *and* $\gamma > 0$. *If* (D) *is feasible, then* z^k *converges. If* (D) *is infeasible, then* z^k *diverges in that* $\|z^k\| \to \infty$.

Proof of Theorem 4.2.13. The qualitative aspect of this theorem (duality between existence of improving directions and strong infeasibility) is known [148]. To the best of our knowledge, the quantitative aspect of this theorem (the meaning and characterization of $P_{\mathcal{N}(A)\cap K}(-c)$) has not been explicitly addressed before. The following proof slightly extends the argument of [148] to show both the qualitative and the quantitative parts.

(P) has no improving direction if and only if

$$\{x \in \mathbb{R}^n | x \in \mathcal{N}(A) \cap K, c^T x < 0\} = \emptyset,$$

which is equivalent to $c^T x \geq 0$ for all $x \in \mathcal{N}(A) \cap K$. This is in turn equivalent to $c \in (\mathcal{N}(A) \cap K)^*$. So

$$-c = P_{-(\mathcal{N}(A)\cap K)^*}(-c).$$

if and only if there is no improving direction, which holds if and only if

$$\mathbf{0} = P_{\mathcal{N}(A)\cap K}(-c).$$

Assume there is an improving direction. Since the projection operator is firmly nonexpansive, we have

$$0 < \|P_{\mathcal{N}(A)\cap K}(-c)\|^2 \leq (P_{\mathcal{N}(A)\cap K}(-c))^T(-c).$$

This simplifies to

$$(P_{\mathcal{N}(A)\cap K}(-c))^T c < 0,$$

and we conclude $P_{\mathcal{N}(A)\cap K}(-c)$ is an improving direction.

Using the fact that $(\mathcal{N}(A) \cap K)^* = \overline{K^* + \mathcal{R}(A^T)}$, we have

$$P_{\mathcal{N}(A)\cap K}(-c) = -P_{\mathcal{N}(A)\cap K}(c) = (P_{\overline{K^*+\mathcal{R}(A^T)}} - I)(c) = P_{\overline{K^*+\mathcal{R}(A^T)-c}}(\mathbf{0}),$$

where we have used the identity $I = P_{\mathcal{N}(A)\cap K} + P_{\overline{K^*+\mathcal{R}(A^T)}}$ in the second equality. □

Proof of Theorem 4.2.14 and 4.2.15. Using the identities $I = P_{\mathcal{N}(A)} + P_{\mathcal{R}(A^T)}$, $I = P_K + P_{-K^*}$, and $R_{\mathcal{R}(A^T)-\gamma c}(z) = R_{\mathcal{R}(A^T)}(z) - 2\gamma Dc$, we have

$$T_3(z) = \tilde{T}(z) - \gamma Dc = \frac{1}{2}(I + R_{\mathcal{R}(A^T)-\gamma c} R_{-K^*})(z).$$

In other words, we can interpret the fixed point iteration

$$z^{k+1} = \tilde{T}(z^k) - \gamma Dc$$

as the DRS iteration on

$$\begin{array}{ll} \text{minimize} & 0 \\ \text{subject to} & x \in \mathcal{R}(A^T) - \gamma c \\ & x \in -K^*. \end{array}$$

This proves Theorem 4.2.15.

Using Lemma 4.2.3, applying Theorem 3.4 of [19] as we did for Theorem 4.2.5, and applying Theorem 4.2.13, we get

$$\begin{aligned} z^k - z^{k+1} &\to P_{\overline{\mathbf{ran}\,(I-T_3)}}(\mathbf{0}) \\ &= P_{\overline{-K^*-\mathcal{R}(A^T)}+\gamma c}(\mathbf{0}) \\ &= -\gamma P_{\overline{K^*+\mathcal{R}(A^T)-c}}(\mathbf{0}) \\ &= -\gamma P_{\mathcal{N}(A)\cap K}(-c). \end{aligned}$$

□

4.2.6 Modifying the objective to achieve finite optimal value

Similar to Theorem 4.2.10, we can achieve strong feasibility of (D) by modifying c, and (P) will have a finite optimal value.

Theorem 4.2.16 (Achieving finite $p^\star$)**.** *Let $w = P_{\overline{K^*+\mathcal{R}(A^T)-c}}(\mathbf{0})$, and let s be any vector satisfying $s \in \mathbf{relint} K^*$. If (P) is feasible and has an unbounded direction, then by replacing c with $c' = c + w + s$, (P) will have a finite optimal value.*

Proof of Theorem 4.2.16. Similar to Lemma 4.2.6, we have

$$w = P_{\overline{P_{\mathcal{N}(A)}(K^*)}-P_{\mathcal{N}(A)(c)}}(\mathbf{0}).$$

And similar to Theorem 4.2.10, the new constraint of (D)

$$K^* \cap \{c + w + s - A^T y\}$$

is strongly feasible. The constraint of (P) is still $K \cap \{x \,|\, Ax = b\}$, which is feasible. By weak duality of we conclude that the optimal value of (P) becomes finite. □

4.2.7 Other cases

So far, we have discussed how to identify and certify cases (a), (d), (f), and (g). We now discuss sufficient conditions to certify the remaining cases.

The following theorem follows from weak duality.

Theorem 4.2.17 ([191] Certificate of finite $p^\star$)**.** *If* (P) *and* (D) *are feasible, then* $p^\star$ *is finite.*

Based on Theorem 4.2.15, we can determine whether (D) is feasible with the iteration $z^{k+1} = T_3(z^k) = \tilde{T}(z^k) - \gamma Dc$,

with any starting point z^0 and $\gamma > 0$.

- $\lim_{k\to\infty} \|z^k\| < \infty$ if and only if (D) is feasible.
- $\lim_{k\to\infty} \|z^k\| = \infty$ if and only if (D) is infeasible.

With a finite number of iterations, we test $\|z^k\| \geq M$ for some large $M > 0$. However, distinguishing the two cases can be numerically difficult as the rate of $\|z^k\| \to \infty$ can be very slow.

Theorem 4.2.18 (Primal iterate convergence)**.** *Consider the DRS iteration as defined in* (4.5) *with any starting point* z^0*. Assume* (P) *is feasible, if* $x^{k+1/2} \to x^\infty$ *and* $x^{k+1} \to x^\infty$*, then* x^∞ *is primal optimal, even if* z^k *doesn't converge.*

When running the fixed-point iteration with $T_1(z) = \tilde{T}(z) + x_0 - \gamma Dc$, if $\|z^k\| \to \infty$ but $x^{k+1/2} \to x^\infty$ and $x^{k+1} \to x^\infty$, then we have case (b), but the converse is not necessarily true.

Proof of Theorem 4.2.18. Define

$$\begin{aligned} x^{k+1/2} &= \mathrm{Prox}_{\gamma g}(z^k) \\ x^{k+1} &= \mathrm{Prox}_{\gamma f}(2x^{k+1/2} - z^k) \\ z^{k+1} &= z^k + x^{k+1} - x^{k+1/2} \end{aligned}$$

as in (4.5) Define

$$\tilde{\nabla} g(x^{k+1/2}) = (1/\gamma)(z^k - x^{k+1/2})$$
$$\tilde{\nabla} f(x^{k+1}) = (1/\gamma)(2x^{k+1/2} - z^k - x^{k+1}).$$

It's simple to verify that

$$\tilde{\nabla} g(x^{k+1/2}) \in \partial g(x^{k+1/2})$$
$$\tilde{\nabla} f(x^{k+1}) \in \partial f(x^{k+1}).$$

Clearly,

$$\tilde{\nabla} g(x^{k+1/2}) + \tilde{\nabla} f(x^{k+1}) = (1/\gamma)(x^{k+1/2} - x^{k+1}).$$

We also have

$$z^{k+1} = z^k - \gamma \tilde{\nabla} g(x^{k+1/2}) - \gamma \tilde{\nabla} f(x^{k+1}) = x^{k+1/2} - \gamma \tilde{\nabla} f(x^{k+1})$$

Consider any $x \in K \cap \{x \,|\, Ax = b\}$. Then, by convexity of f and g,

$$\begin{aligned} g(x^{k+1/2}) - g(x) + f(x^{k+1}) - f(x) &\leq \tilde{\nabla} g(x^{k+1/2})^T (x^{k+1/2} - x) + \tilde{\nabla} f(x^{k+1})^T (x^{k+1} - x) \\ &= (\tilde{\nabla} g(x^{k+1/2}) + \tilde{\nabla} f(x^{k+1}))^T (x^{k+1/2} - x) \\ &\quad + \tilde{\nabla} f(x^{k+1})^T (x^{k+1} - x^{k+1/2}) \\ &= (x^{k+1} - x^{k+1/2})^T (\tilde{\nabla} f(x^{k+1}) - (1/\gamma)(x^{k+1/2} - x)) \\ &= (1/\gamma)(x^{k+1} - x^{k+1/2})^T (x - z^{k+1}) \end{aligned}$$

We take the liminf on both sides and use Lemma 4.2.19 below to get

$$g(x^\infty) + f(x^\infty) \leq g(x) + f(x).$$

Since this holds for any $x \in K \cap \{x \,|\, Ax = b\}$, x^∞ is optimal. □

Lemma 4.2.19. *Let $\Delta^1, \Delta^2, \ldots$ be a sequence in $\mathbb{R}^n$. Then*

$$\liminf_{k \to \infty} (\Delta^k)^T \sum_{i=1}^{k} (-\Delta^i) \leq 0.$$

Proof. Assume for contradiction that

$$\liminf_{k\to\infty}(\Delta^k)^T\sum_{i=1}^{k}(-\Delta^i) > 2\varepsilon$$

for some $\varepsilon > 0$. Since the initial part of the sequence is irrelevant, assume without loss of generality that

$$(\Delta^j)^T\sum_{i=1}^{j}\Delta^i < -\varepsilon$$

for $j = 1, 2, \ldots,$ summing both sides gives us, for all $k = 1, 2, ...$

$$\sum_{j=1}^{k}(\Delta^j)^T\sum_{i=1}^{j}\Delta^i < -\varepsilon k.$$

Define

$$\mathbb{1}\{i \le j\} = \begin{cases} 1, \text{if } i \le j, \\ 0, \text{otherwise.} \end{cases}$$

We have

$$\sum_{j=1}^{k}\sum_{i=1}^{k}(\Delta^j)^T\Delta^i\mathbb{1}\{i \le j\} < -\varepsilon k,$$

$$0 \le \frac{1}{2}\left\|\sum_{i=1}^{k}\Delta^i\right\|^2 + \frac{1}{2}\sum_{i=1}^{k}\left\|\Delta^i\right\|^2 < -\varepsilon k,$$

which is a contradiction. □

4.2.8 The algorithms

We now collect the discussed classification results as three algorithms. The full algorithm is simply running Algorithms 4.1, 4.2, and 4.3, and applying flowchart of Figure 4.1. In theory, the algorithms work with any value of $\gamma > 0$, although the empirical performance can vary with γ.

The algorithms rely on detecting whether certain quantities converge to 0 or ∞. This can be numerically challenging in certain cases. However, certain pathologies are

inherently challenging, and we observe through the examples of Section 4.3 that our method is competitive with other approaches.

Algorithm 4.1 Finding a solution

Parameters: γ, M, ε, z^0

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

$x^{k+1/2} = P_K(z^k)$

$x^{k+1} = D(2x^{k+1/2} - z^k) + x_0 - \gamma Dc$

$z^{k+1} = z^k + x^{k+1} - x^{k+1/2}$

end for

if $\|z^k\| < M$ **then**

Case (a)

$x^{k+1/2}$ and x^{k+1} solution

else if $x^{k+1/2} \to x^\infty$ and $x^{k+1} \to x^\infty$ **then**

Case (b)

$x^{k+1/2}$ and x^{k+1} solution

else

Case (b), (c), (d), (e), (f), or (g).

end if

4.2.9 Case-by-case illustration

In this section, we present a case-by-case illustration of the algorithms. We describe the empirical behavior of the algorithms on cases (b), (c), (d), and (e) and demonstrate how the classification works.

We skip the discussion of case (a), as it is the standard non-pathological case. Algorithm 1 determines whether or not we have case (a). Case (f) and (g) are the infeasible cases, and Algorithm 2 determines whether or not we have case (f) or (g). We skip the discussion of these cases, as we present more thorough experiments of them in

Algorithm 4.2 Feasibility test

Parameters: M, ε, z^0

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

$x^{k+1/2} = P_K(z^k)$

$x^{k+1} = D(2x^{k+1/2} - z^k) + x_0$

$z^{k+1} = z^k + x^{k+1} - x^{k+1/2}$

end for

if $\|z^k\| \geq M$ and $\|z^{k+1} - z^k\| > \varepsilon$ **then**

Case (f)

Strictly separating hyperplane defined by $(z^{k+1} - z^k, ((z^{k+1} - z^k)^T x_0)/2)$

else if $\|z^k\| \geq M$ and $\|z^{k+1} - z^k\| \leq \varepsilon$ **then**

Case (g)

else $\|z^k\| < M$

Case (a), (b), (c), (d), or (e)

end if

Algorithm 4.3 Boundedness test

Prerequisite: (P) is feasible.

Parameters: γ, M, ε, z^0

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

$x^{k+1/2} = P_K(z^k)$

$x^{k+1} = D(2x^{k+1/2} - z^k) - \gamma Dc$

$z^{k+1} = z^k + x^{k+1} - x^{k+1/2}$

end for

if $\|z^k\| \geq M$ and $\|z^{k+1} - z^k\| \geq \varepsilon$ **then**

Case (d)

Improving direction $z^{k+1} - z^k$

else if $\|z^k\| < M$ **then**

Case (a), (b), or (c)

else

Case (a), (b), (c), or (e)

end if

Section 4.3.

Case (b), (P) has a solution but (D) has no solution. Consider the example problem of this case discussed in Section 4.1.2. When we run Algorithm 1, we empirically observe that $\|z^k\| \to \infty$ and $x^{k+1/2}, x^{k+1} \to x^\star$, for $\gamma = 0.1$. This tells us we have case (b).

Case (b), $-\infty < d^\star < p^\star < \infty$. Consider the example problem of this case discussed in Section 4.1.2. When we run Algorithm 1, we empirically observe that $\|z^k\| \to \infty$, $x^{k+1/2}$ and x^{k+1} do not converge, and $\lim_{k\to\infty} 2x_{12}^{k+1} = -0.2$ for $\gamma = 0.1$. When we run Algorithm 2, we empirically observe that z^k converges to a limit. When we run Algorithm 3, we empirically observe that z^k converges to a limit. From this, we can conclude we have case (b) or (c).

Case (b), $-\infty = d^\star < p^\star < \infty$ Consider the problem

$$\begin{array}{ll} \text{minimize} & x_1 \\ \text{subject to} & x_2 - x_3 = 0 \\ & x_3 \geq \sqrt{x_1^2 + x_2^2}, \end{array}$$

which has the solution set $\{(0, t, t) \,|\, t \in \mathbb{R}\}$ and optimal value $p^\star = 0$. Its dual problem is

$$\begin{array}{ll} \text{maximize} & 0 \\ \text{subject to} & y \geq \sqrt{y^2 + 1}, \end{array}$$

which is infeasible. This immediately tells us that $p^\star > -\infty$ is possible even when $d^\star = -\infty$.

We can in fact analyze this example analytically. When we run Algorithm 1 with

starting point $z^0 = (z_1^0, z_2^0, 0)$, the iterates $z^{k+1} = (z_1^{k+1}, z_2^{k+2}, z_3^{k+1})$ are:

$$\begin{aligned} z_1^{k+1} &= \frac{1}{2} z_1^k - \gamma \\ z_2^{k+1} &= \frac{1}{2} z_2^k + \frac{1}{2}\sqrt{(z_1^k)^2 + (z_2^k)^2} \\ z_3^{k+1} &= 0. \end{aligned}$$

So $\|z^k\| \to \infty$. Furthermore, $x^{k+1/2} = P_K(z^k)$ satisfies $x_1^{k+1/2} \to -2\gamma, x_2^{k+1/2} \to \infty$ and $x_3^{k+1/2} \to \infty$, so $x^{k+1/2}$ does not converge to the solution set. When we run Algorithm 2, z^k converges to a limit. When we run Algorithm 3, $\|z^k\| \to \infty$ and $z^{k+1} - z^k \to \mathbf{0}$. From such observations, we could conclude we have case (b), (c), or (e).

This example demonstrates that the converses of Theorem 4.2.17 and 4.2.18 are not true.

Case (c). In this case, $|p^\star| < \infty$ but there is no solution. Consider the example problem of this case discussed in Section 4.1.2. When we run Algorithm 1, we empirically observe that $\|z^k\| \to \infty$, $x^{k+1/2}$ and x^{k+1} do not converge, and $\lim_{k\to\infty} 2x_3^{k+1} = p^\star$ for $\gamma = 0.1$. When we run Algorithm 2, we empirically observe that z^k converges to a limit. When we run Algorithm 3, we empirically observe that z^k converges to a limit. From this, we can conclude we have case (b) or (c).

Case (d). In this case, there is an improving direction. Consider the example problem of this case discussed in Section 4.1.2. When we run Algorithm 1, we empirically observe that $\|z^k\| \to \infty$ and $x^{k+1/2}$, x^{k+1} do not converge for $\gamma = 0.1$. When we run Algorithm 2, we empirically observe that z^k converges to a limit. When we run Algorithm 3, we empirically observe that $\|z^k\| \to \infty$ and $\lim_{k\to\infty} \|z^{k+1} - z^k\| > 0$. From this, we can conclude we have case (d).

Case (e). In this case, $p^\star = -\infty$, but there is no improving direction. Consider the example problem of this case discussed in Section 4.1.2. When we run Algorithm 1,

we empirically observe that $\|z^k\| \to \infty$ and $x^{k+1/2}$, x^{k+1} do not converge for $\gamma = 0.1$. When we run Algorithm 2, we empirically observe that z^k converges to a limit. When we run Algorithm 3, we empirically observe that $\|z^k\| \to \infty$ and $z^{k+1} - z^k \to \mathbf{0}$. From this, we can conclude we have case (b), (c), or (e).

4.3 Numerical Experiments

We test our algorithm on a library of weakly infeasible SDPs generated by [135]. These semidefinite programs are in the form:

$$\begin{aligned} \text{minimize} \quad & C \bullet X \\ \text{subject to} \quad & A_i \bullet X = b_i, i = 1, ..., m \\ & X \in S^n_+, \end{aligned}$$

where $n = 10$, $m = 10$ or 20, and $A \bullet B = \sum_{i=1}^n \sum_{j=1}^n A_{ij} B_{ij}$ denotes the inner product between two $n \times n$ matrices A and B.

The library provides "clean" and "messy" instances. Given a clean instance, a messy instance is created with

$$\begin{aligned} A_i &\leftarrow U^T (\sum_{j=1}^m T_{ij} A_j) U \text{ for } i = 1, ..., m \\ b_i &\leftarrow \sum_{j=1}^m T_{ij} b_j \text{ for } i = 1, ..., m, \end{aligned}$$

where $T \in \mathbb{Z}^{m \times m}$ and $U \in \mathbb{Z}^{n \times n}$ are random invertible matrices with entries in $[-2, 2]$.

In [135], four solvers are tested, specifically, SeDuMi, SDPT3 and MOSEK from the YALMIP environment, and the preprocessing algorithm of Permenter and Parrilo [176] interfaced with SeDuMi. Table 4.1 reports the numbers of instances determined infeasible out of 100 weakly infeasible instances. The four solvers have varying success in detecting infeasibility of the clean instances, but none of them succeed in the messy instances.

Our proposed method performs better. However, it does require many iterations

Table 4.1: Percentage of infeasibility detection in [135], C stands for “clean" and M stands for “messy".

	$m = 10$		$m = 20$	
	C	M	C	M
SeDuMi	0	0	1	0
SDPT3	0	0	0	0
MOSEK	0	0	11	0
PP+SeDuMi	100	0	100	0

Table 4.2: Percentage of infeasibility detection success, C stands for “clean" and M stands for “messy".

	$m = 10$		$m = 20$	
	C	M	C	M
Proposed method	100	21	100	99

Table 4.3: Percentage of success determination that problems are not strongly infeasible, C stands for “clean" and M stands for “messy".

	$m = 10$		$m = 20$	
	C	M	C	M
Proposed method	100	100	100	100

and does fail with some of the messy instances. We run the algorithm with $N = 10^7$ iterations and label an instance infeasible if $1/\|z^N\| \leq 8 \times 10^{-2}$ (cf. Theorem 4.2.7 and 4.2.8). Table 4.2 reports the numbers of instances determined infeasible out of 100 weakly infeasible instances. Curiously, our method and other existing methods perform better with the larger instances of $m = 20$. This behavior is also reported and discussed in [135], the paper that provides the library of pathological instances. We suspect this phenomenon is inherent to the data set, not our algorithm.

We would like to note that detecting whether or not a problem is strongly infeasible is easier than detecting whether a problem is infeasible. With $N = 5 \times 10^4$ and a tolerance of $\|z^N - z^{N+1}\| < 10^{-3}$ (c.f Theorem 4.2.8) our proposed method correctly determined that all test instances are not strongly infeasible. Table 4.3 reports the

numbers of instances determined not strongly infeasible out of 100 weakly infeasible instances.

CHAPTER 5

DRS and ADMM for Pathological Convex Problems

5.1 Introduction

Douglas–Rachford splitting (DRS) and alternating directions method of multipliers (ADMM) are classical methods originally presented in [173, 77, 133, 116] and [94, 103], respectively. DRS and ADMM are closely related. Over the last decade, these methods have enjoyed a resurgence of popularity, as the demand to solve ever larger problems grew.

DRS and ADMM have strong theoretical guarantees and empirical performance, but such results are often limited to non-pathological problems; in particular, most analyses assume a primal solution exists, a dual solution exists, and strong duality holds. When these assumptions are not met, i.e., under pathologies, the theory often breaks down and the empirical performance may degrade significantly. Surprisingly, there had been very little work analyzing DRS and ADMM under pathologies, despite the vast literature on these methods. There has been some recent exciting progress in this area, which we review in Section 5.1.2.

In this chapter, we analyze the asymptotic behavior of DRS and ADMM under pathologies. While it is well known that the iterates "diverge" in such cases, the precise manner in which they do so was not known. We establish that when strong duality holds, i.e., when $p^\star = d^\star \in [-\infty, \infty]$, DRS works, in the sense that asymptotically the divergent iterates are approximately feasible and approximately optimal. The assumption that primal and dual solutions exist is not necessary. We then translate the

pathological analyses for DRS to pathological analyses for ADMM.

Furthermore, we conjecture that DRS necessarily fails when strong duality fails, and we present empirical evidence that supports (but does not prove) this conjecture. In other words, we believe strong duality is the necessary and sufficient condition for DRS to work.

5.1.1 Summary of results, contribution, and organization

Sections 5.4 and 5.5 present what we consider the fruits of this work, the convergence analyses of DRS and ADMM under various pathologies. In fact, we suggest readers read Sections 5.4 and 5.5 before reading the theory of Section 5.3, as doing so will give a sense of direction.

We quickly illustrate, through examples, the kinds of results we show. Precise definitions and statements are presented later. We want DRS and ADMM to find a point that is *approximately feasible* and, when applicable, *approximately optimal*. For example, if the primal problem is weakly infeasible, we want the DRS iterates to satisfy

$$x^{k+1} - x^{k+1/2} \to \mathbf{0}$$

and we show this as Theorem 5.4.6. As another example, if the primal problem is feasible but has no solution and $d^\star = p^\star > -\infty$, we want the DRS iterates to satisfy

$$x^{k+1} - x^{k+1/2} \to \mathbf{0}, \qquad f(x^{k+1/2}) + g(x^{k+1}) \to p^\star$$

and we show related results as Theorems 5.4.3 and 5.4.4. We can say something for all the pathological cases, so long as $d^\star = p^\star$.

Section 5.3 presents the main theoretical contribution of this work. To show that DRS and ADMM successfully achieve the 2 goals of approximate feasibility and approximate optimality, we need 2 separate major theoretical components.

Section 5.3.1 presents the first component, which analyzes the "fixed-point iteration" without a fixed point with tools from operator theory. With this machinery, we show

results like $x^{k+1} - x^k \to \mathbf{0}$ or $x^{k+1} - x^k \to v$, where v is a certificate of (primal or dual) infeasibility. Our contribution is defining the notion of improving directions via recession functions and fully characterizing the infimal displacement vector with this notion.

Section 5.3.2 presents the second component, the function-value analysis, which is based on ideas from convex optimization and subgradient inequalities. With these techniques, we show results like $f(x^{k+1/2})+g(x^{k+1}) \to p^\star$. This part requires the $d^\star = p^\star$ assumption. Our function-value analysis uses, but does not immediately follow from, the results of Section 5.3.1. To the best of our knowledge, analyzing the convergence of objective values for DRS or ADMM applied to pathological problems has not been done before.

Section 5.3.3 presents a third, relatively minor theoretical component, which we use later in Section 5.5 to translate analyses for DRS to analyses for ADMM.

As the goal of this work is to prove several theorems, one each for the many pathological cases, we build up our theory in a series of lemmas and corollaries. Some of these lemmas are rather simple extensions of known results while some are novel. All results of Section 5.3 are eventually used in proving the 5 theorems of Section 5.4 and the 3 theorems of Section 5.5.

The chapter is organized as follows. Section 5.2 reviews standard notions of convex analysis, states several known results, and sets up the notation. Section 5.3 presents the main theoretical contributions. Section 5.4 analyzes DRS under pathologies with the theory of Section 5.3. Section 5.5 analyzes ADMM under pathologies with the theory of Sections 5.4 and 5.3. Section 5.6 presents counterexamples to make additional observations. Section 5.7 concludes this chapter.

5.1.2 Prior work

As pathological convex optimization problems do arise in practice [142, 71, 78, 225, 229], there is practical value in studying how well-behaved and robust an algorithm is in such setups However, the there had been surprisingly little work investigating the behavior of the popular methods DRS and ADMM under pathologies. The understanding is still incomplete, but there has been some recent progress: [19, 24, 27, 137] analyze DRS under specific pathological setups, [22, 23, 27] analyze DRS under general setups, and [183, 210, 12] analyze ADMM under specific pathological setups for conic programs. These studies, however, are limited to more specific setups and pathologies where an improving direction exists or the primal problem is strongly feasible.

The convex feasibility problem of finding an $x \in A \cap B$, where A and B are nonempty closed convex sets, is a subclass of problems with practical importance. While it is possible to recast convex feasibility problems into equivalent optimization problems and apply the results of this work, prior work on the specific setup has stronger results [19, 24, 27]. We discuss further comparisons in Section 5.4.4.

DRS has strong primal-dual symmetry, in the sense of Fenchel duality for convex optimization [91, 190] and, more generally, Attouch-Théra duality for monotone operators [152, p. 40] and [8]. See [80, Lemma 3.6 p. 133] or [13, 26, 27] for in-depth studies on this subject. Naturally, our results also exhibit a degree of primal-dual symmetry, although we do not explicitly address it in the interest of space. Rather, we take the viewpoint that the primal problem is the problem of interest and the dual problem is an auxiliary conceptual and computational tool.

In operator theory, and especially in infinite dimensional problems arising from physics and PDEs, the sum of two maximal monotone operators may not be maximal, and one can consider this a pathology. One remedy to such pathology is to generalize the notion of the sum of two operators by regularizing the operators and then considering the limit as the regularization is reduced to zero [9, 186, 187]. This notion of pathology

and the remedy is quite different from what we consider. For some of the pathologies we consider, $\partial f + \partial g$ is a perfectly well-defined maximal monotone operator. Moreover, we do not remedy the pathology but rather simply analyze how DRS and ADMM behave under the pathology. We work in finite dimensions and thereby avoid the notion of weak and strong convergence.

When a problem is known to be pathological a priori, one can first modify or regularize the problem and then solve the non-pathological problem. One such approach is facial reduction, a pre-processing step that rids a pathological conic program of difficult pathologies [37, 39, 35, 184, 170, 55, 226, 228, 178, 144, 175, 177, 245]. In contrast, the goal of this work is to analyze DRS and ADMM when directly applied to pathological convex programs. To put in differently, we do not assume users of DRS or ADMM have a priori knowledge of whether the problem is pathological.

The standard analysis for DRS proves the iterates converge using ideas from operator theory and fixed point iterations [133, 80, 81, 57, 59, 85, 58]. The standard analyses of ADMM prove the iterates converge by reducing ADMM to DRS [95, 81, 85] or with a direct analysis via a Lyapunov function [92, 102, 32, 74, 53]. These analyses rely on the existence of a primal-dual saddle point, which only exists under the non-pathological case, and therefore do not immediately generalize to pathological setups.

The first part of our analysis relies on a classical result by Pazy [172] and Baillon et al. [11] from the 1970s, which characterize the asymptotic behavior of fixed-point iterations without fixed-points. There has been some recent work that analyze algorithms that can be interpreted as fixed-point iterations without fixed-points [14, 19, 40, 6, 22, 20, 24, 7, 158, 27, 198, 137]. The analysis of Section 5.3.1 was inspired by these works.

Another recent line of analysis for DRS and ADMM is function-value analysis, which establishes the objective values, rather than the iterates, converge [65, 67, 68]. These analyses, however, also rely on the existence of a primal-dual saddle point and do not immediately generalize to the pathological setups. The function-value analysis of

Section 5.3.2 was inspired by these works.

5.2 Preliminaries

In addition to the preliminaries introduced in Sec. 1.4, we also need the following in the chapter. Define the recession function of a PCC function f as

$$\operatorname{rec} f(d) = \lim_{\alpha \to \infty} \frac{f(x + \alpha d) - f(x)}{\alpha} \tag{5.1}$$

for any $x \in \mathbf{dom}\, f$. Loosely speaking, the recession function characterizes the asymptotic change of f as we go in direction d. In fact,

$$f(x + \alpha d) = \alpha \operatorname{rec} f(d) + o(\alpha)$$

as $\alpha \to \infty$ for any $x \in \mathbf{dom}\, f$. The recession function $\operatorname{rec} f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a positively homogeneous PCC function. If $h(x) = g(-x)$, then $\operatorname{rec}(h^*)(d) = \operatorname{rec}(g^*)(-d)$. When f and g are PCC, either $f(x)+g(x) = \infty$ for all $x \in \mathbb{R}^n$ or $\operatorname{rec}(f+g) = \operatorname{rec} f + \operatorname{rec} g$. If f is PCC, then $\sigma_{\mathbf{dom}\, f^*} = \operatorname{rec} f$.

Define the proximal operator $\operatorname{Prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ as

$$\operatorname{Prox}_f(z) = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left\{ f(x) + (1/2)\|x - z\|^2 \right\}.$$

When f is PCC, the arg min uniquely exists, and therefore Prox_f is well-defined. When C is closed and convex, $\operatorname{Prox}_{\delta_C} = \Pi_C$. When f is PCC, $\operatorname{Prox}_f + \operatorname{Prox}_{f^*} = I$, where $I : \mathbb{R}^n \to \mathbb{R}^n$ is the identity operator.

A mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive if $\|T(x) - T(y)\| \leq \|x - y\|$ for all $x, y \in \mathbb{R}^n$. Nonexpansive mappings are, by definition, Lipschitz continuous with Lipschitz constant 1. $T : \mathbb{R}^n \to \mathbb{R}^n$ is firmly-nonexpansive if

$$\|T(x) - T(y)\|^2 \leq \langle x - y, T(x) - T(y) \rangle$$

for all $x, y \in \mathbb{R}^n$. Proximal and projection operators are firmly-nonexpansive.

5.2.1 Duality and primal subvalue

We call the optimization problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x)+g(x), \tag{P}$$

the primal problem. We call the optimization problem

$$\underset{\nu\in\mathbb{R}^n}{\text{maximize}} \quad -f^*(\nu)-g^*(-\nu), \tag{D}$$

the dual problem. Throughout this chapter, we assume f and g are PCC.

(P) is feasible if $\mathbf{0}\in\mathbf{dom}\, f-\mathbf{dom}\, g$, strongly infeasible if $\mathbf{0}\notin\overline{\mathbf{dom}\, f-\mathbf{dom}\, g}$, and weakly infeasible otherwise. (P) falls under exactly one of the three cases. (P) is infeasible if it is not feasible. (D) is feasible if $\mathbf{0}\in\mathbf{dom}\,(f^*)+\mathbf{dom}\,(g^*)$, strongly infeasible if $\mathbf{0}\notin\overline{\mathbf{dom}\,(f^*)+\mathbf{dom}\,(g^*)}$, and weakly infeasible otherwise.

We call $p^\star=\inf\{f(x)+g(x)\,|\,x\in\mathbb{R}^n\}$ the primal optimal value and $d^\star=\sup\{-f^*(\nu)-g^*(-\nu)\,|\,\nu\in\mathbb{R}^n\}$ the dual optimal value. We let $p^\star=\infty$ if (P) is infeasible and $d^\star=-\infty$ if (D) is infeasible. Weak duality, which always holds, states $d^\star\leq p^\star$. We say strong duality holds between (P) and (D), if $d^\star=p^\star\in[-\infty,\infty]$. We say *total duality* holds between (P) and (D), if (P) has a solution, (D) has a solution, and strong duality holds.

Define the primal subvalue of (P) as

$$p^-=\lim_{\varepsilon\to 0^+}\inf_{x,y\in\mathbb{R}^n}\{f(x)+g(y)\,|\,\|x-y\|\leq\varepsilon\}.$$

The notion of primal subvalue is standard in conic programming [118, 234, 149, 147]. Here, we generalize it to general convex programs. The following theorem is well known [195], although we have not seen it stated exactly in this form.

Theorem 5.2.1. *If f and g are PCC, then $d^\star=p^-\leq p^\star$.*

In fact, the following proof follows the exposition of [195].

Proof. Write $h(\nu) = f^*(\nu) + g^*(-\nu)$, and define

$$p(\delta) = \min_{x \in \mathbb{R}^n} \{f(x + \delta) + g(x)\}.$$

Since f and g are proper, i.e., finite somewhere, p is proper. Since p is defined by partial minimization of a convex function, it is convex.

Then

$$\begin{aligned}
p^*(\nu) &= -\min_{\delta \in \mathbb{R}^n} \{p(\delta) - \nu^T \delta\} \\
&= -\min_{\delta \in \mathbb{R}^n} \left\{ \min_{x \in \mathbb{R}^n} \{f(x + \delta) + g(x)\} - \nu^T \delta \right\} \\
&= -\min_{x \in \mathbb{R}^n} \left\{ \min_{\delta \in \mathbb{R}^n} \{f(x + \delta) - \nu^T \delta\} + g(x) \right\} \\
&= -\min_{x \in \mathbb{R}^n} \left\{ \min_{\delta' \in \mathbb{R}^n} \{f(\delta') - \nu^T \delta'\} + \nu^T x + g(x) \right\} \\
&= f^*(\nu) - \min_{x \in \mathbb{R}^n} \left\{ \nu^T x + g(x) \right\} \\
&= f^*(\nu) + g^*(-\nu) = h(\nu).
\end{aligned}$$

We can rewrite the definition of the primal subvalue as

$$p^- = \lim_{\varepsilon \to 0} \inf_{\|\delta\| \leq \varepsilon} p(\delta) = \liminf_{\delta \to \mathbf{0}} p(\delta),$$

where the second equality follows from the definition of lim inf. The lower semi-continuous hull of p is p^{**} [195, Theorem 4 and 5], i.e.,

$$\liminf_{\delta \to \mathbf{0}} p(\delta) = p^{**}(\mathbf{0}).$$

So

$$p^- = p^{**}(\mathbf{0}) = h^*(\mathbf{0}) = \sup_{\nu \in \mathbb{R}^n} \{f^*(\nu) + g^*(-\nu)\} = d^*.$$

□

With Theorem 5.2.1, we can interpret strong duality as well-posedness of (P). The primal subvalue p^- is the optimal value of (P) achieved with infinitesimal infeasibilities. When the infinitesimal infeasibilities provide a non-infinitestimal improvement to the function value, we can consider (P) ill-posed.

5.2.2 Douglas–Rachford operator

Douglas–Rachford splitting (DRS) applied to (P) is

$$\begin{aligned} x^{k+1/2} &= \operatorname{Prox}_{\gamma f}(z^k) \\ x^{k+1} &= \operatorname{Prox}_{\gamma g}(2x^{k+1/2} - z^k) \\ z^{k+1} &= z^k + x^{k+1} - x^{k+1/2} \end{aligned} \tag{5.2}$$

with a starting point $z^0 \in \mathbb{R}^n$ and a parameter $\gamma > 0$. We also express this iteration more concisely as $z^{k+1} = T_\gamma(z^k)$ where

$$T_\gamma = \frac{1}{2}I + \frac{1}{2}(2\operatorname{Prox}_{\gamma g} - I)(2\operatorname{Prox}_{\gamma f} - I).$$

$T_\gamma : \mathbb{R}^n \to \mathbb{R}^n$ is a firmly-nonexpansive operator, and we interpret DRS as a fixed-point iteration. Write T_1 for T_γ with $\gamma = 1$.

The standard analysis of DRS assumes total duality, which, again, means (P) has a solution, (D) has a solution, and $d^\star = p^\star$.

Theorem 5.2.2 (Theorem 7.1 and 8.1 of [13] and Proposition 4.8 of [80])**.** *Total duality holds between* (P) *and* (D) *if and only if T_γ has a fixed point for some $\gamma > 0$. If total duality holds between* (P) *and* (D)*, then DRS converges in that $z^k \to z^\star$, where $x^\star = \operatorname{Prox}_{\gamma f}(z^\star)$ is a solution of* (P)*. If total duality does not hold between* (P) *and* (D)*, then DRS diverges in that $\|z^k\| \to \infty$.*

Theorem 5.2.2 is well known, although the term "total duality" is not always used. More often, total duality is assumed by instead assuming a saddle point exists for an appropriate Lagrangian.

5.2.3 Fixed-point iterations without fixed points

Theorem 5.2.2 states the DRS iteration has no fixed points under pathologies. Analyzing fixed-point iterations without fixed points is the first part of our pathological analysis.

Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a firmly-nonexpansive operator. Write

$$\mathbf{ran}\,(I - T) = \{z - T(z) \mid z \in \mathbb{R}^n\}.$$

Note that T has a fixed point if and only if $\mathbf{0} \in \mathbf{ran}\,(I - T)$. The closure of this set, $\overline{\mathbf{ran}\,(I - T)}$, is closed and convex [172]. We call

$$v = \Pi_{\overline{\mathbf{ran}\,(I-T)}}(\mathbf{0})$$

the *infimal displacement vector* of T. (The term was coined in [22].) If T has a fixed point, then $v = \mathbf{0}$, but $v = \mathbf{0}$ is possible even when T has no fixed point.

The following classical result by Pazy and Baillon et al. elegantly characterizes the asymptotic behavior of fixed-point iterations with respect to T.

Theorem 5.2.3 (Theorem 2 of [172] and Corollary 2.3 of [11])**.** *If T is firmly-nonexpansive and v is its infimal displacement vector, the iteration $z^{k+1} = T(z^k)$ satisfies*

$$z^k = -kv + o(k), \qquad z^{k+1} - z^k \to -v.$$

Theorem 5.2.3 is especially powerful when we can concretely characterize v. Recently, Bauschke, Hare, and Moursi published the following elegant formula.

Theorem 5.2.4 ([23])**.** *The infimal displacement vector v of T_1, the DRS operator, satisfies*

$$v = \arg\min \left\{ \|z\| \mid z \in \overline{\mathbf{dom}\, f - \mathbf{dom}\, g} \cap \overline{\mathbf{dom}\, f^* + \mathbf{dom}\, g^*} \right\}.$$

The original result in [23] is more general as it applies to the DRS operator of monotone operators. In Section 5.3.1, we use Theorem 5.2.4 and the notion of improving directions to provide a further concrete characterization of v.

5.3 Theoretical results

In this section, we present the main theoretical contribution of this chapter. Our analysis requires a generalized notion of improving directions, so we define it first. Section 5.3.1 analyzes DRS as a fixed-point iteration without fixed points. Section 5.3.2

analyzes DRS as an optimization method that reduces function values. Section 5.3.3 directly analyzes the evolution of the $x^{k+1/2}$ and x^{k+1}-iterates of DRS. Later in Sections 5.4 and 5.5 we combine these results to analyze the asymptotic behavior of DRS and ADMM applied to pathological convex programs.

While the formula of Theorem 5.2.4 is known, the use of improving directions to concretely characterize the infimal displacement vector v is new. An improving direction may or may not exist, and we analyze both cases. Our analysis shows that existence of an improving direction is a key deciding factor in how DRS behaves.

We say $d \in \mathbb{R}^n$ is a *primal improving direction* for (P) if

$$\operatorname{rec} f(d) + \operatorname{rec} g(d) < 0.$$

Note $\operatorname{rec} f(d) + \operatorname{rec} g(d) = \operatorname{rec}(f+g)(d)$ when (P) is feasible. For simplicity, we only consider primal improving directions when (P) is feasible. The notion of (primal) improving direction is standard in conic programming [149, 162, 147]. Here, we extend it to general convex programs of the form (P).

If (P) is feasible and there is a primal improving direction, then $p^\star = -\infty$. To see why, let d be a primal improving direction. Then

$$f(x+\alpha d) + g(x+\alpha d) = \alpha \operatorname{rec}(f+g)(d) + o(\alpha)$$

for any $x \in \mathbf{dom}\, f \cap \mathbf{dom}\, g$ as $\alpha \to \infty$, and therefore $p^\star = -\infty$. However, $p^\star = -\infty$ is possible even when (P) has no improving direction. We discuss such an example in Section 5.4.

Likewise, we say $d' \in \mathbb{R}^n$ is a *dual improving direction* if

$$\operatorname{rec}(f^*)(d') + \operatorname{rec}(g^*)(-d') < 0.$$

If (D) is feasible and there is a dual improving direction, then $d^\star = \infty$.

5.3.1 Infimal displacement vector of the DRS operator

In this section, we provide a further concrete characterization of the infimal displacement vector v. When (P) or (D) is strongly infeasible, Theorem 5.2.4 states $v \neq \mathbf{0}$. Our contribution is to show v is an improving direction in this case. For the sake of simplicity, we first analyze T_1 and then translate the results to T_γ for $\gamma > 0$.

We first consider the case where (P) is feasible and characterize v based on the primal improving direction or the absence of it.

Lemma 5.3.1. (P) *has an improving direction if and only if* (D) *is strongly infeasible. Write*

$$d = -\Pi_{\overline{(\mathbf{dom}\, f^* + \mathbf{dom}\, g^*)}}(\mathbf{0}).$$

If (P) *has an improving direction, then* $d \neq \mathbf{0}$ *and* d *is an improving direction. If* (P) *has no improving direction, then* $d = \mathbf{0}$.

Proof. We first show

$$-\Pi_{\overline{(\mathbf{dom}\, f^* + \mathbf{dom}\, g^*)}}(\mathbf{0}) = \mathrm{Prox}_{\mathrm{rec}\, f + \mathrm{rec}\, g}(\mathbf{0}). \tag{5.3}$$

Let A and B be nonempty convex sets. The identities of Section 5.2 tell us

$$(\delta_{\overline{A+B}})^*(x) = \sigma_{\overline{A+B}}(x) = \sigma_{A+B}(x) = \sigma_A(x) + \sigma_B(x).$$

Setting $A = \mathbf{dom}\, f^*$ and $B = \mathbf{dom}\, g^*$ gives us

$$(\delta_{\overline{\mathbf{dom}\, f^* + \mathbf{dom}\, g^*}})^*(x) = \sigma_{\mathbf{dom}\, f^*}(x) + \sigma_{\mathbf{dom}\, g^*}(x).$$

Based on the identities of Section 5.2, we have

$$\begin{aligned}
\Pi_{\overline{\mathbf{dom}\, f^* + \mathbf{dom}\, g^*}}(\mathbf{0}) &= \mathrm{Prox}_{\delta_{\overline{\mathbf{dom}\, f^* + \mathbf{dom}\, g^*}}}(\mathbf{0}) \\
&= (I - \mathrm{Prox}_{\sigma_{\overline{\mathbf{dom}\, f^* + \mathbf{dom}\, g^*}}})(\mathbf{0}) \\
&= -\mathrm{Prox}_{\sigma_{\mathbf{dom}\, f^* + \mathbf{dom}\, g^*}}(\mathbf{0}) \\
&= -\mathrm{Prox}_{\sigma_{\mathbf{dom}\, f^*} + \sigma_{\mathbf{dom}\, g^*}}(\mathbf{0}) \\
&= -\mathrm{Prox}_{\mathrm{rec}\, f + \mathrm{rec}\, g}(\mathbf{0}).
\end{aligned}$$

Remember that $\operatorname{rec} f + \operatorname{rec} g$ is a convex positively homogeneous function. Since $\operatorname{rec} f(\mathbf{0}) + \operatorname{rec} g(\mathbf{0}) = 0$,

$$\mathbf{0} = \arg\min \left\{ \operatorname{rec} f(x) + \operatorname{rec} g(x) + (1/2)\|x\|^2 \right\} = \operatorname{Prox}_{\operatorname{rec} f + \operatorname{rec} g}(\mathbf{0})$$

if and only if $\operatorname{rec} f(x) + \operatorname{rec} g(x) \geq 0$ for all $x \in \mathbb{R}^n$. By our definition of an improving direction, $\operatorname{rec} f(x) + \operatorname{rec} g(x) \geq 0$ for all $x \in \mathbb{R}^n$ if and only if there is no improving direction. By definition, $\mathbf{0} = \Pi_{\overline{\mathbf{dom}\, f^* + \mathbf{dom}\, g^*}}(\mathbf{0})$ if and only if (D) is not strongly infeasible. So with (5.3), we conclude (P) has an improving direction if and only if (D) is strongly infeasible.

It remains to show that

$$d = \arg\min \left\{ \operatorname{rec} f(x) + \operatorname{rec} g(x) + (1/2)\|x\|^2 \right\}$$

is an improving direction, if $d \neq \mathbf{0}$. Since d is defined as a minimizer, we have

$$\operatorname{rec} f(d) + \operatorname{rec} g(d) + (1/2)\|d\|^2 \leq \operatorname{rec} f(\mathbf{0}) + \operatorname{rec} g(\mathbf{0}) + (1/2)\|\mathbf{0}\|^2 = 0.$$

This implies $\operatorname{rec} f(d) + \operatorname{rec} g(d) \leq -(1/2)\|d\|^2 < 0$, i.e., d is an improving direction. □

Lemma 5.3.2. *Assume* (P) *is feasible. Then*

$$v = -d = \Pi_{\overline{\mathbf{dom}\, f^* + \mathbf{dom}\, g^*}}(\mathbf{0})$$

is the infimal displacement vector of T_1.

Proof. Let x_0 be a feasible point of (P). Since $\operatorname{rec} f(d) + \operatorname{rec} g(d) \leq 0 < \infty$ by Lemma 5.3.1 and the definition of an improving direction, we have $x_0 \in \mathbf{dom}\, f$, $x_0 + d \in \mathbf{dom}\, g$, and thus $-d \in \mathbf{dom}\, f - \mathbf{dom}\, g \subseteq \overline{\mathbf{dom}\, f - \mathbf{dom}\, g}$. Since $-d$ is the minimum-norm element of $\overline{\mathbf{dom}\, f^* + \mathbf{dom}\, g^*}$, Theorem 5.2.4 tells us that $-d$ is the infimal displacement vector of T_1. □

Corollary 5.3.3. *Assume* (P) *is feasible, and* (D) *is feasible. Then* $v = \mathbf{0}$ *is the infimal displacement vector of* T_γ *for any* $\gamma > 0$.

Corollary 5.3.4. *Assume* (P) *is feasible, and* (D) *is weakly infeasible. Then* $v = \mathbf{0}$ *is the infimal displacement vector of* T_γ *for any* $\gamma > 0$.

Corollary 5.3.5. *Assume* (P) *is feasible, and* (D) *is strongly infeasible. Then*

$$v = -\gamma d = \gamma \Pi_{\overline{\mathbf{dom}\, f^* + \mathbf{dom}\, g^*}}(\mathbf{0}) \neq \mathbf{0}$$

is the infimal displacement vector of T_γ *for any* $\gamma > 0$. *Furthermore,* d *is an improving direction of* (P).

Next, we consider the case where (D) is feasible and characterize the infimal displacement vector based on the dual improving direction or the absence of it.

Lemma 5.3.6. *Assume* (D) *is feasible. Then*

$$v = -d' = \Pi_{\overline{\mathbf{dom}\, f - \mathbf{dom}\, g}}(\mathbf{0})$$

is the infimal displacement vector of T_1.

Proof. Following the same logic as in the proof of Lemma 5.3.1, we have

$$\Pi_{\overline{\mathbf{dom}\, f - \mathbf{dom}\, g}}(\mathbf{0}) = -\arg\min_{\nu} \left\{ \mathrm{rec}(f^*)(\nu) + \mathrm{rec}(g^*)(-\nu) + (1/2)\|\nu\|^2 \right\},$$

and

$$d' = -\Pi_{\overline{\mathbf{dom}\, f - \mathbf{dom}\, g}}(\mathbf{0})$$

is a dual improving direction, if $d' \neq \mathbf{0}$.

Let ν_0 be any feasible point of (D). Then $\nu_0 \in \mathbf{dom}\, f^*$ and $-\nu_0 - d' \in \mathbf{dom}\, g^*$. Therefore, $-d' \in \mathbf{dom}\, f^* + \mathbf{dom}\, g^* \subseteq \overline{\mathbf{dom}\, f^* + \mathbf{dom}\, g^*}$. Since $-d'$ is defined to be the minimum-norm element of $\overline{\mathbf{dom}\, f - \mathbf{dom}\, g}$ we conclude the statement with Theorem 5.2.4. □

Corollary 5.3.7. *Assume* (D) *is feasible, and* (P) *is weakly infeasible. Then* $v = \mathbf{0}$ *is the infimal displacement vector of* T_γ *for any* $\gamma > 0$.

Corollary 5.3.8. *Assume* (D) *is feasible, and* (P) *is strongly infeasible. Then*

$$v = -d' = \Pi_{\overline{\mathbf{dom}\, f - \mathbf{dom}\, g}}(\mathbf{0}) \neq \mathbf{0},$$

is the infimal displacement vector of T_γ *for any* $\gamma > 0$. *Furthermore,* d' *is a dual improving direction.*

Note that for Corollary 5.3.8, the infimal displacement vector is independent of the value of γ.

5.3.2 Function-value analysis

In this section, we present the second major theoretical component to our analysis. Section 5.3.1 analyzed the infimal displacement vector of T_γ. This, however, is not sufficient for characterizing the asymptotic behavior of DRS in relation to the original optimization problem (P).

Let us briefly discuss why function-value analysis is necessary. Consider the convex function $h(x, y) = x^2/y$ defined for $y > 0$. Note that h has minimizers, $(0, y)$ for any $y > 0$, and the operator $I - \nabla h$ has fixed points. It is straightforward to verify that $h(\sqrt{y}, y) - \inf f \nrightarrow 0$, but $\nabla h(\sqrt{y}, y) \to 0$ as $y \to \infty$, i.e., $(\sqrt{y}, y)$ for large y is not an approximate minimizer for h but does approximate satisfy the fixed point condition for $I - \nabla h$. It is possible to construct a similar example with the DRS operator. If we let $f = h$ and $g = 0$, then DRS reduces to the proximal point method on h. This operator exhibits the same exact issue.

This means **approximate fixed points do not always correspond to approximate solutions of the original problem**. This is why we need a separate and distinct function-value analysis to accompany the fixed-point theory.

We now present function-value analysis. Throughout this section, write $x^{k+1/2}$ and x^{k+1} to denote the DRS iterates of (5.2).

Lemma 5.3.9. *For all $k = 0, 1, \ldots$ and any $x \in \mathbb{R}^n$*

$$f(x^{k+1/2}) + g(x^{k+1}) - f(x) - g(x) \\ \leq (1/\gamma)\langle x^{k+1} - x^{k+1/2}, x - z^{k+1}\rangle.$$

An inequality similar to that of Lemma 5.3.9 has been presented as Proposition 2 of [67]. We nevertheless quickly show a direct proof.

Proof. Write

$$\tilde{\nabla} f(x^{k+1/2}) = (1/\gamma)(z^k - x^{k+1/2}) \\ \tilde{\nabla} g(x^{k+1}) = (1/\gamma)(2x^{k+1/2} - z^k - x^{k+1}).$$

From the definition of the DRS iteration (5.2), we can verify that

$$\tilde{\nabla} f(x^{k+1/2}) \in \partial f(x^{k+1/2}), \qquad \tilde{\nabla} g(x^{k+1}) \in \partial g(x^{k+1})$$

and that

$$\tilde{\nabla} f(x^{k+1/2}) + \tilde{\nabla} g(x^{k+1}) = (1/\gamma)(x^{k+1/2} - x^{k+1}).$$

We also have

$$z^{k+1} = z^k - \gamma\tilde{\nabla} f(x^{k+1/2}) - \gamma\tilde{\nabla} g(x^{k+1}) = x^{k+1/2} - \gamma\tilde{\nabla} g(x^{k+1}).$$

If $x \notin \mathbf{dom}\, f \cap \mathbf{dom}\, g$, then $f(x) + g(x) = \infty$ for all $x \in \mathbb{R}^n$, and there is nothing to prove. Now, consider any $x \in \mathbf{dom}\, f \cap \mathbf{dom}\, g$. Then, by definition of subdifferentials,

$$\begin{aligned} &f(x^{k+1/2}) - f(x) + g(x^{k+1}) - g(x) \\ &\quad \leq \langle \tilde{\nabla} f(x^{k+1/2}), x^{k+1/2} - x\rangle + \langle \tilde{\nabla} g(x^{k+1}), x^{k+1} - x\rangle \\ &\quad = (\tilde{\nabla} f(x^{k+1/2}) + \langle \tilde{\nabla} g(x^{k+1}), x^{k+1/2} - x\rangle + \langle \tilde{\nabla} g(x^{k+1}), x^{k+1} - x^{k+1/2}\rangle \\ &\quad = \langle x^{k+1} - x^{k+1/2}, \tilde{\nabla} g(x^{k+1}) - (1/\gamma)(x^{k+1/2} - x)\rangle \\ &\quad = (1/\gamma)\langle x^{k+1} - x^{k+1/2}, x - z^{k+1}\rangle. \end{aligned}$$

□

The following result, which is well known for non-pathological setups, also holds under pathologies, so long as $d^\star = p^\star$.

Lemma 5.3.10. *Assume $p^\star = d^\star \in [-\infty, \infty]$. Assume the infimal displacement vector v of T_γ satisfies $v = \mathbf{0}$. Then*

$$\lim_{k\to\infty} \frac{1}{k+1} \sum_{i=0}^{k} f(x^{i+1/2}) + g(x^{i+1}) = p^\star$$

and

$$\liminf_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) = p^\star.$$

Proof. If $\Delta^0, \Delta^1, \dots$ is any sequence in $\mathbb{R}^n$, then

$$\begin{aligned}\sum_{j=0}^{k} \langle \Delta^j, \sum_{i=0}^{j} \Delta^i \rangle &= \sum_{j=0}^{k}\sum_{i=0}^{k} \mathbb{1}\{i \le j\} \langle \Delta^j, \Delta^i \rangle \\ &= \frac{1}{2}\left\| \sum_{i=0}^{k} \Delta^i \right\|^2 + \frac{1}{2} \sum_{i=0}^{k} \left\| \Delta^i \right\|^2.\end{aligned}$$

Let $\Delta^k = z^{k+1} - z^k = x^{k+1} - x^{k+1/2}$ and sum the inequality of Lemma 5.3.9 to get

$$\begin{aligned}\gamma \sum_{i=0}^{k} f(x^{i+1/2}) - f(x) + g(x^{i+1}) - g(x) &\le \sum_{j=0}^{k} \langle \Delta^j, x - z^0 \rangle - \sum_{j=0}^{k} \langle \Delta^j, \sum_{i=0}^{j} \Delta^i \rangle \\ &= \langle z^{k+1} - z^0, x - z^0 \rangle - \frac{1}{2}\|z^{k+1} - z^0\|^2 - \frac{1}{2}\sum_{i=0}^{k} \|z^{i+1} - z^i\|^2 \\ &= -\frac{1}{2}\|z^{k+1}\|^2 + \frac{1}{2}\|z^0\|^2 + \langle z^{k+1} - z^0, x \rangle - \frac{1}{2}\sum_{i=0}^{k} \|z^{i+1} - z^i\|^2.\end{aligned}$$

Divide both sides by $(k+1)/2$ to get

$$\begin{aligned}&\frac{2\gamma}{k+1} \sum_{i=0}^{k} \Big(f(x^{i+1/2}) - f(x) + g(x^{i+1}) - g(x) \Big) \\ &\le -\frac{1}{k+1}\|z^{k+1}\|^2 + \frac{1}{k+1}\|z^0\|^2 - \frac{1}{k+1}\sum_{i=0}^{k} \|z^{i+1} - z^i\|^2 + \frac{2}{k+1}\langle z^{k+1} - z^0, x \rangle.\end{aligned} \tag{5.4}$$

for all $k = 0, 1, \dots$ and any $x \in \mathbb{R}^n$.

We now show

$$\limsup_{k\to\infty} \frac{1}{k+1} \sum_{i=0}^{k} \Big(f(x^{i+1/2}) + g(x^{i+1}) \Big) \le p^\star.$$

Assume $p^\star < \infty$, as otherwise there is nothing to prove. Let x be any $x \in \mathbf{dom}\, f \cap \mathbf{dom}\, g$. By Theorem 5.2.3, $z^k = -kv + o(k)$. If $v \neq \mathbf{0}$, then the first (negative) term on the right-hand side of (5.4) dominates the positive terms. If $v = \mathbf{0}$, then both nonnegative terms on the right-hand side of (5.4) converge to 0. In both cases, we have

$$\limsup_{k\to\infty} \frac{1}{k+1} \sum_{i=0}^{k} f(x^{i+1/2}) + g(x^{i+1}) \leq f(x) + g(x) \tag{5.5}$$

for all $x \in \mathbf{dom}\, f \cap \mathbf{dom}\, g$. We minimize the right-hand side to obtain $p^\star$.

By Theorem 5.2.3, $v = \mathbf{0}$ implies $x^{k+1/2} - x^{k+1} \to \mathbf{0}$. In turn, by Theorem 5.2.1, we have

$$\liminf_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) \geq p^\star.$$

Combining this with (5.5) gives us the first stated result.

It is straightforward to verify that if a real-valued sequence a^k satisfies

$$\liminf_{k\to\infty} a^k \geq a, \qquad \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} a^i = a,$$

then

$$\liminf_{k\to\infty} a^k = a.$$

The second stated result follows from this argument. □

Lemma 5.3.10 provides the function-value analysis when $v = \mathbf{0}$, and the first part of Lemma 5.3.11 provides the analysis when $v \neq \mathbf{0}$. The later parts part of Lemma 5.3.11 is used in translating the analyses for DRS to analyses for ADMM in Section 5.5.

Lemma 5.3.11. *Assume* (P) *is feasible and* $v \neq \mathbf{0}$*, i.e.,* (P) *has an improving direction. Then*

$$f(x^{k+1/2}) + g(x^{k+1}) \to p^\star = -\infty.$$

Moreover, $|f(x^{k+1/2})| \leq \mathcal{O}(k)$ *and* $|g(x^{k+1})| \leq \mathcal{O}(k)$ *as* $k \to \infty$*. Assume* (P) *is feasible and* $v = \mathbf{0}$*. Then* $|f(x^{k+1/2})| \leq o(k)$ *and* $|g(x^{k+1})| \leq o(k)$ *as* $k \to \infty$*.*

Proof. When (P) has an improving direction, Corollary 5.3.5 and Theorem 5.2.3 tells us

$$z^{k+1} - z^k = x^{k+1} - x^{k+1/2} \to \gamma d.$$

Then Lemma 5.3.9 tells us that

$$(1/k)(f(x^{k+1/2}) + g(x^{k+1})) \leq -(1/\gamma)\langle x^{k+1} - x^{k+1/2}, (1/k)z^{k+1}\rangle + O(1/k)$$

which tells us

$$\limsup_{k\to\infty}(1/k)(f(x^{k+1/2}) + g(x^{k+1})) \leq -\gamma\|d\|^2. \tag{5.6}$$

This proves the first statement.

Assume $v = \mathbf{0}$. With the same reasoning as for (5.6) we get

$$\limsup_{k\to\infty}(1/k)(f(x^{k+1/2}) + g(x^{k+1})) \leq 0.$$

Assume (P) feasible, without making any asumptions on v. Write $\tilde{\nabla} f(x^{1/2})$ for any subgradient of f at $x^{1/2}$. Then

$$\begin{aligned} f(x^{k+1/2}) &\geq f(x^{1/2}) + \langle \tilde{\nabla} f(x^{1/2}), x^{k+1/2} - x^{1/2}\rangle \\ &\geq f(x^{1/2}) - \|\tilde{\nabla} f(x^{1/2})\|\|x^{k+1/2} - x^{1/2}\| = k\gamma\|d\|\|\tilde{\nabla} f(x^{1/2})\| + o(k), \end{aligned}$$

and we conclude

$$\liminf_{k\to\infty}(1/k)f(x^{k+1/2}) \geq -\gamma\|d\|\|\tilde{\nabla} f(x^{1/2})\|.$$

With a similar argument, we get

$$\liminf_{k\to\infty}(1/k)g(x^{k+1}) \geq -\gamma\|d\|\|\tilde{\nabla} g(x^1)\|$$

where $\tilde{\nabla} g(x^1)$ is any subgradient of g at x^1. Combining these with (5.6) gives us the remaining statements. □

Lemma 5.3.12. *Assume $p^\star = d^\star$. Assume $x^{k+1/2}$ and x^{k+1} are the DRS iterates as defined in (5.2). If $x^{k+1/2}, x^{k+1} \to x^\star$ for some $x^\star$, then $x^\star$ is a solution.*

Proof. We first note that closed functions are by definition lower semi-continuous, and that f and g are assumed to be closed. By Lemma 5.3.10 we have

$$f(x^\star) + g(x^\star) \leq \liminf_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) = p^\star,$$

and we conclude $f(x^\star) + g(x^\star) = p^\star$. □

5.3.3 Evolution of shadow iterates

Section 5.3.1 characterized the evolution of the z^k-iterates, which we could call the main iterates. The $x^{k+1/2}$ and x^{k+1}-iterates of DRS are called the *shadow iterates*. Here, we analyze the evolution of the shadow iterates.

Although the results of this section are are not as fundamental or important as the results of Sections 5.3.1 and 5.3.2, we do need these results later, especially when translating the analyses for DRS to analyses for ADMM.

Lemma 5.3.13. *If $v = \mathbf{0}$, then $x^{k+3/2} - x^{k+1/2} \to \mathbf{0}$ and $x^{k+2} - x^{k+1} \to \mathbf{0}$.*

Proof. Since $v = \mathbf{0}$, we have $z^{k+1} - z^k \to \mathbf{0}$. Since the map the defines $z^k \mapsto x^{k+1/2}$ and $z^{k+1} \mapsto x^{k+3/2}$ is Lipschitz continuous, $x^{k+3/2} - x^{k+1/2} \to \mathbf{0}$. Finally, $z^{k+1} - z^k \to \mathbf{0}$ and $x^{k+3/2} - x^{k+1/2} \to \mathbf{0}$ implies $x^{k+2} - x^{k+1} \to \mathbf{0}$. □

Lemma 5.3.14. *If (P) is strongly infeasible and (D) is feasible, then $x^{k+3/2} - x^{k+1/2} \to \mathbf{0}$ and $x^{k+2} - x^{k+1} \to \mathbf{0}$.*

Proof. Write $-d'$ for the infimal displacement vector as given by Corollary 5.3.8. By Theorem 5.2.3, we have

$$z^{k+1} - z^k = x^{k+1} - x^{k+1/2} \to d'.$$

The projection inequality states

$$\langle v - \Pi_C x, \Pi_C x - x \rangle \geq 0 \tag{5.7}$$

for any nonempty closed convex set C, $v \in C$, and $x \in \mathbb{R}^n$. Since $-d' = \Pi_{\overline{\mathbf{dom}\, f - \mathbf{dom}\, g}}(\mathbf{0})$, (5.7) tells us that

$$\langle d', x - x^{k+1} \rangle + \|d'\|^2 \leq 0$$

for any $x \in \mathbf{dom}\, f$. Using $x^{k+1/2} = \mathrm{Prox}_{\gamma f}(z^k)$ and firm-nonexpansiveness of Prox, we get

$$\begin{aligned} \|\mathrm{Prox}_{\gamma f}(z^k + d') - x^{k+1/2}\|^2 &\leq \langle d', \mathrm{Prox}_{\gamma f}(z^k + d') - x^{k+1/2} \rangle \\ &= \langle d', \mathrm{Prox}_{\gamma f}(z^k + d') - x^{k+1} \rangle + \langle d', x^{k+1} - x^{k+1/2} \rangle \\ &\to 0 \end{aligned}$$

since $\langle d', x^{k+1} - x^{k+1/2} \rangle \to \|d'\|^2$. So $\mathrm{Prox}_{\gamma f}(z^k + d') - \mathrm{Prox}_{\gamma f}(z^k) \to \mathbf{0}$. Since Prox is Lipschitz continuous, $z^{k+1} - z^k - d' \to \mathbf{0}$ implies

$$\mathrm{Prox}_{\gamma f}(z^k + d') - \mathrm{Prox}_{\gamma f}(z^{k+1}) \to \mathbf{0}.$$

Putting everything together we conclude

$$\mathrm{Prox}_{\gamma f}(z^{k+1}) - \mathrm{Prox}_{\gamma f}(z^k) = x^{k+3/2} - x^{k+1/2} \to \mathbf{0}.$$

Since

$$z^{k+2} - z^{k+1} = \underbrace{z^{k+1} - z^k}_{\to d'} + x^{k+2} - x^{k+1} - \underbrace{(x^{k+3/2} - x^{k+1/2})}_{\to \mathbf{0}} \to d'$$

we also conclude that $x^{k+2} - x^{k+1} \to \mathbf{0}$. □

Lemma 5.3.15. *If (P) has an improving direction, and (P) is feasible, then $x^{k+3/2} - x^{k+1/2} \to \gamma d$ and $x^{k+2} - x^{k+1} \to \gamma d$, where $-\gamma d = \gamma \Pi_{\overline{\mathbf{dom}\, f^* + \mathbf{dom}\, g^*}}(\mathbf{0})$ is the infimal displacement vector as given in Corollary 5.3.5.*

Proof. For simplicity, assume $\gamma = 1$. For $\gamma \neq 1$, we scale f and g to get the stated result.

Rewrite the DRS iteration as

$$\begin{aligned}
x^{k+1/2} &= \mathrm{Prox}_f(z^k) \\
\nu^{k+1/2} &= z^k - x^{k+1/2} = \mathrm{Prox}_{f^*}(z^k) \\
x^{k+1} &= \mathrm{Prox}_g(2x^{k+1/2} - z^k) \\
\nu^{k+1} &= 2x^{k+1/2} - z^k - x^{k+1} = \mathrm{Prox}_{g^*}(2x^{k+1/2} - z^k) \\
z^{k+1} &= z^k - (\nu^{k+1} + \nu^{k+1/2}).
\end{aligned}$$

By Theorem 5.2.3, we have

$$z^{k+1} - z^k = x^{k+1} - x^{k+1/2} \to d.$$

By the same reasoning as in Lemma 5.3.14, we can use (5.7) and firm-nonexpansiveness to show that

$$\nu^{k+3/2} - \nu^{k+1/2} = \mathrm{Prox}_{f^*}(z^{k+1}) - \mathrm{Prox}_{f^*}(z^k) \to \mathbf{0}.$$

Since

$$z^{k+1} - z^k = \underbrace{\nu^{k+3/2} - \nu^{k+1/2}}_{\to \mathbf{0}} + x^{k+3/2} - x^{k+1/2} \to d,$$

we have $x^{k+3/2} - x^{k+1/2} \to d$.

Since

$$z^{k+2} - z^{k+1} = \underbrace{z^{k+1} - z^k}_{\to d} + x^{k+2} - x^{k+1} - \underbrace{(x^{k+3/2} - x^{k+1/2})}_{\to d} \to d$$

we also conclude that $x^{k+2} - x^{k+1} \to d$. □

5.4 Pathological convergence: DRS

In this section, we use the theory of Section 5.3 to analyze DRS under pathologies. We classify the status of (P) and (D) into 7 cases and provide convergence analyses for the first 6 cases, the ones that assume strong duality.

5.4.1 Classification

The primal-dual problem pair, (P) and (D), falls under exactly one of the following 7 distinct cases.

Case (a) Total duality holds between (P) and (D).

In other words, (P) and (D) have solutions, and $d^\star = p^\star$. For example, the primal problem

$$\text{minimize} \quad x - \log x$$

and its dual problem

$$\begin{aligned} \text{maximize} \quad & 1 + \log(y) \\ \text{subject to} \quad & y = 1 \end{aligned}$$

both have solutions, and $d^\star = p^\star = 1$.

Case (b) $d^\star = p^\star$ is finite, (P) has a solution, (D) has no solution.

For example, the primal problem

$$\text{minimize} \quad \underbrace{\delta_{\{(x_1,x_2)\,|\,x_1^2+x_2^2\leq 1\}}(x_1,x_2)}_{f(x)} + \underbrace{x_2 + \delta_{\{(x_1,x_2)\,|\,x_1=1\}}(x_1,x_2)}_{g(x)}$$

has a solution but its dual problem

$$\text{maximize} \quad -\sqrt{\nu_1^2+\nu_2^2} + \nu_1 - \delta_{\{\nu_2=1\}}(-\nu_2)$$

does not. Nevertheless, $d^\star = p^\star = 0$.

Case (c) $d^\star = p^\star$ is finite, (P) is feasible, but (P) has no solution.

To get such an example, swap the role of the primal and the dual in the example for case (b).

Case (d) $d^\star = p^\star = -\infty$, (P) is feasible, but there is no improving direction.

This implies (D) is weakly infeasible. For example, the primal problem

$$\text{minimize} \quad \delta_{\{x \,|\, x \geq 1\}}(x) - \log x$$

has no solution and has optimal value $p^\star = -\infty$. Since the derivative of the objective, $-1/x$, goes to 0 as $x \to \infty$, the primal problem has no improving direction. The dual problem

$$\begin{array}{ll} \text{maximize} & y + 1 + \log(y) \\ \text{subject to} & y \leq 0 \end{array}$$

is weakly infeasible.

Case (e) $d^\star = p^\star = -\infty$, (P) is feasible, and there is an improving direction.

This implies (D) is strongly infeasible. For example, the primal problem

$$\text{minimize} \quad x + x$$

has an improving direction, namely $d = -1$, and the dual problem

$$\text{maximize} \quad \delta_{\{1\}}(x) + \delta_{\{1\}}(-x)$$

is strongly infeasible.

Case (f) $d^\star = p^\star = \infty$ and (P) is infeasible.

For example, the problem

$$\text{minimize} \quad 1/\sqrt{-x} - \log(x)$$

is infeasible, and its dual

$$\begin{array}{ll} \text{maximize} & (3/2^{2/3})y^{1/3} + 1 + \log(y) \\ \text{subject to} & y \geq 0 \end{array}$$

has optimal value $d^\star = \infty$.

Case (g) $d^\star < p^\star$, i.e. strong duality fails.

5.4.2 Convergence results

Theorem 5.4.1. *[133, 67] In case (a),* $x^{k+1/2}, x^{k+1} \to x^\star$*, where* $x^\star$ *is a solution of* (P) *and*

$$\lim_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) = p^\star.$$

Theorem 5.4.2. *In case (b),* $x^{k+1} - x^{k+1/2} \to \mathbf{0}$ *and*

$$\lim_{k\to\infty} \frac{1}{k+1}\sum_{i=0}^{k} f(x^{i+1/2}) + g(x^{i+1}) = p^\star, \qquad \liminf_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) = p^\star.$$

Furthermore, if $x^{k+1/2} \to x^\star$ *(or equivalently if* $x^{k+1} \to x^\star$*) then* $x^\star$ *is a solution.*

Proof. This follows from Theorem 5.2.3, Corollary 5.3.3, Lemma 5.3.10, and Lemma 5.3.12. □

Theorem 5.4.3. *In case (c),* $x^{k+1} - x^{k+1/2} \to \mathbf{0}$*,*

$$\lim_{k\to\infty} \frac{1}{k+1}\sum_{i=0}^{k} f(x^{i+1/2}) + g(x^{i+1}) = p^\star, \qquad \liminf_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) = p^\star,$$

and $(x^{k+1/2}, x^{k+1})$ *do not converge.*

Proof. This follows from Theorem 5.2.3, Corollary 5.3.3, Lemma 5.3.10, and the contrapositive of Lemma 5.3.12. □

Theorem 5.4.4. *In case (d), (D) is weakly infeasible,* $x^{k+1} - x^{k+1/2} \to \mathbf{0}$*,*

$$\lim_{k\to\infty} \frac{1}{k+1}\sum_{i=0}^{k} f(x^{i+1/2}) + g(x^{i+1}) = -\infty, \qquad \liminf_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) = -\infty,$$

and $(x^{k+1/2}, x^{k+1})$ *do not converge.*

Proof. This follows from Theorem 5.2.3, Lemma 5.3.1, Corollary 5.3.4, Lemma 5.3.10, and the contrapositive of Lemma 5.3.12. □

Theorem 5.4.5. *In case (e), (D) is strongly infeasible, $x^{k+1} - x^{k+1/2} \to \gamma d$, where d is an improving direction,*

$$\lim_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) = -\infty,$$

and $(x^{k+1/2}, x^{k+1})$ do not converge. Furthermore, $\operatorname{dist}(x^{k+1/2}, \mathbf{dom}\, g) \to 0$ *and* $\operatorname{dist}(x^{k+1}, \mathbf{dom}\, f) \to 0$.

Proof. All but the last assertions follows from Theorem 5.2.3, Lemma 5.3.1, Corollary 5.3.5, Lemma 5.3.11, and the contrapositive of Lemma 5.3.12. By Lemma 5.3.15 $x^{k+1/2} - x^{k-1/2} \to \gamma d$ and by Theorem 5.2.3 and Corollary 5.3.5 $x^k - x^{k-1/2} \to \gamma d$. So $x^{k+1/2} - x^k \to \mathbf{0}$. Since $x^k \in \mathbf{dom}\, g$, we have

$$\operatorname{dist}(x^{k+1/2}, \mathbf{dom}\, g) \leq \operatorname{dist}(x^{k+1/2}, x^k) \to 0.$$

Since $x^{k+1/2} \in \mathbf{dom}\, f$, we have

$$\operatorname{dist}(x^k, \mathbf{dom}\, f) \leq \operatorname{dist}(x^k, x^{k+1/2}) \to 0.$$

□

Theorem 5.4.6. *In case (f),* $\|x^{k+1} - x^{k+1/2}\| \to \operatorname{dist}(\mathbf{dom}\, f, \mathbf{dom}\, g)$.

Proof. This follows from Theorem 5.2.3 and Corollaries 5.3.7 and 5.3.8. □

5.4.3 Interpretation

We can view the DRS as an algorithm with two major goals: make the iterates feasible and optimal. With some caveats, DRS succeeds at both. As an auxiliary goal, we want the shadow iterates of DRS to converge to a solution if one exists. With some caveats, DRS succeeds at this as well. Finally, DRS provides a certificate of infeasibility in cases (e) and (f).

In cases (a), (b), (c), and (d) the iterates become approximately feasible in that $x^{k+1} - x^{k+1/2} \to \mathbf{0}$. In case (e) the iterates become approximately feasible in that

$\operatorname{dist}(x^{k+1/2}, \mathbf{dom}\, g) \to 0$ and $\operatorname{dist}(x^{k+1}, \mathbf{dom}\, f) \to 0$. In case (f), feasibility is impossible, but DRS does its best to achieve feasibility.

In cases (a), (b), (c), (d), and (e), the function values on average converge to the optimal value. In other words, DRS finds the correct optimal value in these cases.

In case (a), the shadow iterates, the $x^{k+1/2}$ and x^{k+1} iterates, converge to a solution. In case (b), we do not know whether the shadow iterates converge to a solution. However, if they converge, the limit is a solution. In cases (c), (d), and (e), the shadow iterates do not converge, which is good since there is no solution to converge to.

In cases (e) and (f), the limit $z^{k+1} - z^k \to -v \neq \mathbf{0}$ provides a certificate of dual and primal strong infeasibility, respectively. These may be computationally useful when verifying the validity of a certificate is easy, which is the case for conic programs.

We quickly clarify the contribution. The analysis of case (a) is well known and is not the focus of this work, but we include it's discussion here for completeness. Approximate feasibility in cases (a), (b), (c) and (d) directly follows from prior work, in particular from Theorems 5.2.3 and 5.2.4. The approximate feasibility results for cases (e) and (f) are contributions of this work.

5.4.4 Feasibility problems

Consider the problem of finding an $x \in A \cap B$, where A and B are nonempty closed convex sets. Recasting this convex feasibility problem into an equivalent optimization problem and using Theorem 5.2.4 [23], Theorem 5.2.3 [172, 11], Theorem 5.4.1 [133], and basic convex analysis provides us the following results:

- Case (a). If $A \cap B \neq \emptyset$ then $x^{k+1/2}, x^{k+1} \to x^\star$ where $x^\star \in A \cap B$.
- Case (f). If $\operatorname{dist}(A, B) > 0$, then $\|x^{k+1} - x^k\| \to \operatorname{dist}(A, B)$.
- Case (g). If $A \cap B \neq \emptyset$ but $\operatorname{dist}(A, B) = 0$, then $x^{k+1/2} - x^{k+1} \to 0$.

Specifically, one can recast the convex feasibility problem $x \in A \cap B$ into the primal problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \delta_A(x) + \delta_B(x),$$

which has the dual problem

$$\underset{\nu \in \mathbb{R}^n}{\text{maximize}} \quad -\sigma_A(\nu) - \sigma_B(-\nu).$$

When $A \cap B \neq \emptyset$, then $p^\star = 0$ with $x \in A \cap B$ and $d^\star = 0$ with $\nu = 0$. Therefore total duality holds (i.e., we have case (a)) and Theorem 5.4.1 applies. When $\operatorname{dist}(A, B) > 0$, then $p^\star = \infty$ since $A \cap B = \emptyset$. For the dual, define $\tilde{\nu} = P_{\overline{A-B}}(0)$, which satisfies

$$\langle a - b, \tilde{\nu} \rangle \geq \|\tilde{\nu}\|^2$$

for all $a \in A$ and $b \in B$ by the optimality conditions defining the projection. Then we have

$$-\sigma_A(-\eta\tilde{\nu}) - \sigma_B(+\eta\tilde{\nu}) = \inf_{a \in A, b \in B} \langle a - b, \tilde{\nu} \rangle \geq \eta \|\tilde{\nu}\|^2$$

for $\eta > 0$. Since $\|\tilde{\nu}\| = \operatorname{dist}(A, B) > 0$, with $\eta \to \infty$ we conclude $d^\star = \infty$. So we have case (f) and Theorem 5.4.6 applies. However, the results of this work say nothing for case (g). The contribution of this work is to consider improving directions and function-value analysis, but both notions are not relevant in the setup of convex feasibility problems. Therefore, our work does not provide any new results for the convex feasibility problems.

Prior work on the convex feasibility setup provides further stronger results. By [19, Theorem 3.13], we have

$$x^{k+1} - x^{k+1/2} \to \Pi_{\overline{B-A}}(\mathbf{0}).$$

Furthermore, by [27, Theorem 4.5], we have

$$(x^{k+1/2}, x^{k+1}) \to (a^{\text{apx}}, b^{\text{apx}}) \in \underset{(a,b) \in A \times B}{\arg\min} \{\|a - b\|\}$$

if the arg min is nonempty. (The pairs in the arg min are called "best approximation pairs" between A and B.) These results show that the relevant dichotomy is whether a

best approximation pair exists, rather than whether strong duality holds. These results cannot be obtained from the analysis of our work.

5.5 Pathological convergence: ADMM

We now analyze ADMM under pathologies. Consider the primal problem

$$\begin{array}{ll} \underset{x\in\mathbb{R}^p, y\in\mathbb{R}^q}{\text{minimize}} & f(x)+g(y) \\ \text{subject to} & Ax+By=c, \end{array} \tag{P-ADMM}$$

where $f:\mathbb{R}^p\to\mathbb{R}\cup\{\infty\}$ and $g:\mathbb{R}^q\to\mathbb{R}\cup\{\infty\}$ are PCC, $A\in\mathbb{R}^{n\times p}$, $B\in\mathbb{R}^{n\times q}$, and $c\in\mathbb{R}^n$, and its dual problem

$$\underset{\nu\in\mathbb{R}^n}{\text{maximize}} \quad -f^*(-A^T\nu)-g^*(-B^T\nu)-c^T\nu. \tag{D-ADMM}$$

Write $p^\star$ and $d^\star$ for the primal and dual optimal values. ADMM applied to this primal-dual problem pair is

$$\begin{aligned} x^{k+1} &\in \underset{x\in\mathbb{R}^p}{\arg\min}\left\{f(x)+\langle\nu^k, Ax+By^k-c\rangle+\frac{1}{2\gamma}\|Ax+By^k-c\|^2\right\} \\ y^{k+1} &\in \underset{y\in\mathbb{R}^q}{\arg\min}\left\{g(y)+\langle\nu^k, Ax^{k+1}+By-c\rangle+\frac{1}{2\gamma}\|Ax^{k+1}+By-c\|^2\right\} \\ \nu^{k+1} &= \nu^k+(1/\gamma)(Ax^{k+1}+By^{k+1}-c). \end{aligned} \tag{5.8}$$

For ADMM to be well-defined, the argmins of (5.8) must exist. Throughout this section, we furthermore assume the regularity conditions

$$(\mathbf{ran}\, A^T)\cap \operatorname{ri}\mathbf{dom}\,(f^*)\neq\emptyset, \tag{5.9}$$

$$(\mathbf{ran}\, B^T)\cap \operatorname{ri}\mathbf{dom}\,(g^*)\neq\emptyset. \tag{5.10}$$

Here, ri denotes the relative interior of a set. These conditions ensure the subproblems are solvable [190, Theorem 16.3].

Without these regularity conditions, the subproblems of (5.8) may not have solutions. This is often overlooked and sometimes even misunderstood throughout the

ADMM literature. (The highly influential paper [41] mistakenly claimed it is enough for f and g to be PCC. Chen, Sun, and Toh [53] pointed out that additional assumptions are needed.)

5.5.1 Classification and convergence results

Under (5.9) and (5.10), the status of (P-ADMM) and (D-ADMM) falls under exactly one of the following 5 distinct cases.

Case (a) $d^\star = p^\star$, both (P-ADMM) and (D-ADMM) have solutions.

Theorem 5.5.1 ([103, 41, 67])**.** *In case (a),* $Ax^k + By^k - c \to \mathbf{0}$ *and*

$$\lim_{k\to\infty} f(x^k) + g(y^k) \to p^\star.$$

Case (b) $d^\star = p^\star$, (P-ADMM) has a solution, (D-ADMM) has no solution.

Theorem 5.5.2. *In case (b),* $Ax^k + By^k - c \to \mathbf{0}$ *and*

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} f(x^i) + g(y^i) = p^\star, \quad \liminf_{k\to\infty} f(x^k) + g(y^k) = p^\star.$$

Furthermore, if $(x^k, y^k) \to (x^\star, y^\star)$*, then* $(x^\star, y^\star)$ *is a solution.*

Case (c) $d^\star = p^\star \in [-\infty, \infty)$, (P-ADMM) is feasible but has no solution.

Theorem 5.5.3. *In case (c),* $Ax^k + By^k - c \to \mathbf{0}$ *and*

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} f(x^i) + g(y^i) = p^\star, \quad \liminf_{k\to\infty} f(x^k) + g(y^k) = p^\star,$$

and the sequence (x^k, y^k) *does not converge.*

Case (d) $d^\star = p^\star = \infty$, (P-ADMM) is infeasible.

Theorem 5.5.4. *In case (d),*

$$\|Ax^k + By^k - c\| \to \inf_{\substack{x\in\mathbf{dom}\, f\\ y\in\mathbf{dom}\, g}} \|Ax + By - c\|.$$

Case (e) $d^\star < p^\star$, i.e. strong duality fails.

5.5.2 Interpretation

With some caveats, ADMM succeeds at achieving feasibility and optimality. In cases (a), (b), and (c) the iterates become approximately feasible in that $Ax^k + By^k - c \to \mathbf{0}$, and the function values on average converge to the solution. In case (d), feasibility is impossible, but ADMM does its best to achieve feasibility.

5.5.3 Proofs

ADMM is often analyzed as DRS applied to (D-ADMM) [95]. In this proof, however, we take the less common approach shown in [80, 238], which derives ADMM directly from the primal problem. We do so as the function-value analysis of Section 5.3.2 translate nicely with this primal approach.

Consider the equivalent primal optimization problem

$$\underset{z \in \mathbb{R}^n}{\text{minimize}} \quad \tilde{f}(z) + \tilde{g}(z)$$

with

$$\tilde{f}(z) = \inf\{f(x) \,|\, Ax + z = \mathbf{0}\}, \quad \tilde{g}(z) = \inf\{g(y) \,|\, By - c = z\},$$

which are PCC functions, as we assume (5.9) and (5.10) [190, Theorem 16.3]. We apply DRS to this form to get

$$\begin{aligned}
\tilde{x}^{k+1/2} &= \arg\min_{\tilde{x}} \left\{ \gamma \tilde{g}(\tilde{x}) + (1/2)\|\tilde{x} - z^k\|^2 \right\} \\
\tilde{x}^{k+1} &= \arg\min_{\tilde{x}} \left\{ \gamma \tilde{f}(\tilde{x}) + (1/2)\|\tilde{x} - 2\tilde{x}^{k+1/2} + z^k\|^2 \right\} \\
z^{k+1} &= z^k + \tilde{x}^{k+1} - \tilde{x}^{k+1/2},
\end{aligned}$$

where we perform the $\tilde{g}$-update before the $\tilde{f}$-update. We introduce and substitute the variables x^k, y^k, and ν^k defined implicitly by $\tilde{x}^{k+1/2} = By^{k+1} - c$, $\tilde{x}^{k+1} = -Ax^{k+2}$, and

$z^k = -\gamma\nu^k - Ax^{k+1}$ to get

$$
\begin{aligned}
y^{k+1} &= \arg\min_y \left\{\gamma g(y) + \gamma\langle \nu^k, Ax^{k+1} + By - c\rangle + (1/2)\|Ax^{k+1} + By - c\|^2\right\} \\
x^{k+2} &= \arg\min_x \left\{\gamma f(x) + \gamma\langle \nu^{k+1}, Ax + By^{k+1} - c\rangle + (1/2)\|Ax + By^{k+1} - c\|^2\right\} \\
\nu^{k+1} &= \nu^k + (1/\gamma)(Ax^{k+1} + By^{k+1} - c).
\end{aligned}
$$

Reordering the updates to get the dependency right, we get

$$
\begin{aligned}
y^{k+1} &= \arg\min_y \left\{\gamma g(y) + \gamma\langle \nu^k, Ax^{k+1} + By - c\rangle + (1/2)\|Ax^{k+1} + By - c\|^2\right\} \\
\nu^{k+1} &= \nu^k + (1/\gamma)(Ax^{k+1} + By^{k+1} - c) \\
x^{k+2} &= \arg\min_x \left\{\gamma f(x) + \gamma\langle \nu^{k+1}, Ax + By^{k+1} - c\rangle + (1/2)\|Ax + By^{k+1} - c\|^2\right\}.
\end{aligned}
$$

Finally, redefine the start and end of an iteration so that it updates x^{k+1}, y^{k+1}, and ν^{k+1} instead y^{k+1}, ν^{k+1}, and x^{k+2}. With this, we get (5.8).

The the last step, where we redefine the start and end of an iteration, introduces a subtlety when translating the results of Section 5.4.2. In particular, the results of Section 5.3.3 are necessary because of this.

Theorem 5.5.2 follows from Theorem 5.4.2 and Lemmas 5.3.2, 5.3.11, and 5.3.13. Theorem 5.5.4 follows from Theorem 5.4.6 and Lemma 5.3.14.

Case (c) of this section corresponds to cases (c), (d), and (e) of Section 5.4.2. For the three cases, we use Theorem 5.4.3 and Lemmas 5.3.2, 5.3.11, and 5.3.13, Theorem 5.4.4 and Lemmas 5.3.2, 5.3.11, and 5.3.13, and Theorem 5.4.5 and Lemma 5.3.11, and 5.3.15. Combining the three results into one gives us Theorem 5.5.3. □

5.6 When strong duality fails

In the analyses of DRS, we assumed strong duality holds. When strong duality fails, i.e., when $d^\star < p^\star$, we conjecture that DRS fails.

Conjecture. *When strong duality fails, DRS necessarily fails in that*

$$\liminf_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) < p^\star.$$

In other words, DRS finds the wrong objective value.

As discussed in Section 5.4.2, DRS tries to achieve feasibility and optimality. As discussed in Section 5.2.1, strong duality is well-posedness. Therefore, when the problem is ill-posed, we expect DRS to reduce the function value below $p^\star$ while achieving an infinitesimal infeasibility. We support the conjecture with examples.

We first present an analytical counter example. Consider the problem taken from [137]

$$\text{minimize} \quad \underbrace{\delta_{\{(x_1,x_2,x_3)\,|\,x_3\geq(x_1^2+x_2^2)^{1/2}\}}(x)}_{f(x)} + \underbrace{x_1 + \delta_{\{(x_1,x_2,x_3)\,|\,x_2=x_3\}}(x)}_{g(x)}$$

which has the solution set $\{(0,t,t)\,|\,t\in\mathbb{R}\}$ and optimal value $p^\star = 0$. Its dual problem

$$\text{maximize} \quad -\delta_{\{(\nu_1,\nu_2,\nu_3)\,|\,-\nu_3\geq(\nu_1^2+\nu_2^2)^{1/2}\}}(\nu) - \delta_{\{(\nu_1,\nu_2,\nu_3)\,|\,\nu_1=1,\,\nu_2=-\nu_3\}}(-\nu)$$

is infeasible. Given $z^0 = (z_1^0, z_2^0, 0)$, the DRS iterates have the form

$$\begin{aligned} z_1^{k+1} &= \frac{1}{2}z_1^k - \gamma \\ z_2^{k+1} &= \frac{1}{2}z_2^k + \frac{1}{2}\sqrt{(z_1^k)^2 + (z_2^k)^2} \\ z_3^{k+1} &= 0. \end{aligned}$$

With this, it is relatively straightforward to show $x^{k+1/2} - x^{k+1} \to \mathbf{0}$, $x_1^{k+1/2} \to -2\gamma$, $x_2^{k+1/2} \to \infty$, $x_3^{k+1/2} \to \infty$, and $f(x^{k+1/2}) + g(x^{k+1}) \to -2\gamma$. Also, $x^{k+1/2} \nrightarrow \mathbf{dom}\, f \cap \mathbf{dom}\, g$ even though $x^{k+1/2} - x^{k+1} \to \mathbf{0}$.

Note that

$$d^\star < \lim_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) < p^\star.$$

So this counterexample proves, at least in some cases, that DRS solves neither the primal nor the dual problem in the absence of strong duality.

Next, we present more experimental counter examples that support the conjecture. We run DRS on these problems report the experimental results.

The problem, taken from [31],

$$\underset{x\in\mathbb{R}^2}{\text{minimize}} \quad \underbrace{\exp(-\sqrt{x_1x_2})}_{f(x)} + \underbrace{\delta_{\{(x_1,x_2)\,|\,x_1=0\}}(x)}_{g(x)}$$

has $p^\star = 1$ but $d^\star = 0$. Experimentally, for all $\gamma > 0$ and choice of z^0 we observe $d^\star < \lim_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) < p^\star$.

The problem, taken from [78],

$$\underset{X\in\mathbf{S}^3}{\text{minimize}} \quad \underbrace{\delta_{\mathbf{S}^3_+}(X)}_{f(X)} + \underbrace{X_{22} + \delta_{\{X\in\mathbf{S}^3\,|\,X_{33}=0, X_{22}+2X_{13}=1\}}(X)}_{g(X)},$$

where $\mathbf{S}^3$ and $\mathbf{S}^3_+$ respectively denote the set of symmetric and positive semidefinite 3×3 matrices, has $p^\star = 1$ but $d^\star = 0$. Experimentally, we observe $d^\star = \lim_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1})$ for $\gamma \geq 0.5$, and $d^\star < \lim_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) < p^\star$ for $0 < \gamma < 0.5$. This behavior does not depend on z^0.

The problem, taken from [239],

$$\underset{X\in\mathbf{S}^3}{\text{minimize}} \quad \underbrace{\delta_{\mathbf{S}^3_+}(X)}_{f(X)} + \underbrace{2X_{12} + \delta_{\{X\in\mathbf{S}^3\,|\,X_{22}=0, -2X_{12}+2X_{33}=2\}}(X)}_{g(X)}$$

has $p^\star = 0$ but $d^\star = -2$. Experimentally, we observe $d^\star = \lim_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1})$ for $\gamma \geq 1$, and $d^\star < \lim_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) < p^\star$ for $0 < \gamma < 1$. This behavior does not depend on z^0.

The problem, taken from [219],

$$\underset{X\in\mathbf{S}^5}{\text{minimize}} \quad \underbrace{\delta_{\mathbf{S}^5_+}(X)}_{f(X)} + \underbrace{X_{44} + X_{55} + \delta_{\{X\in\mathbf{S}^3\,|\,X_{11}=0, X_{22}=1, X_{34}=1, 2X_{13}+2X_{45}+X_{55}=1\}}(X)}_{g(X)}$$

has $p^\star = (\sqrt{5}-1)/2$ but $d^\star = 0$. Experimentally, we observe $d^\star = \lim_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1})$ for $\gamma \geq 0.8$, and $d^\star < \lim_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) < p^\star$ for $0 < \gamma < 0.8$. This behavior does not depend on z^0.

The conjecture holds for all examples. Interestingly, for some examples, there is a threshold $\gamma_{\min}$ such that $d^\star < \lim_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1}) < p^\star$ when $0 < \gamma < \gamma_{\min}$ and $d^\star = \lim_{k\to\infty} f(x^{k+1/2}) + g(x^{k+1})$ when $\gamma_{\min} \leq \gamma$. We do not have an explanation for this phenomenon.

5.7 Conclusion

In this chapter, we analyzed DRS and ADMM under pathologies. We show that when strong duality holds, the iterates of DRS and ADMM are approximately feasible and approximately optimal in the sense discussed in Sections 5.4.3 and 5.5.2. Furthermore, we conjectured that DRS necessarily fails when strong duality fails, and we provided empirical evidence supporting this conjecture.

As discussed in Section 5.6, DRS exhibits an interesting behavior in the absence of strong duality, and we do not have an explanation for it. Analyzing this phenomenon and addressing the conjecture is an interesting direction of future research.

For non-pathological problems, DRS can be generalized with an over-under relaxation parameter between 0 and 2. The pathological DRS analysis of this chapter immediately extends to this generalized setup. For non-pathological problems, ADMM can be generalized with an over-under relaxation parameter between 0 and $(1+\sqrt{5})/2$. This generalization arises when ADMM is analyzed directly through a Lyapunov function, and not through DRS [92, 102, 32, 89, 74, 53, 52]. The pathological ADMM analysis of this chapter *does not* immediately extend to this generalized setup. Analyzing this form of ADMM applied to pathological problems is an interesting direction of future research.

Part IV

Convergence Behaviors on Nonconvex Problems

In this part, we present the results of [139], in which the convergence behaviors of FBS and DRS on nonconvex problems are analyzed. The main result is that under some smoothness conditions, FBS and DRS can avoid the strict saddle points[1] almost surely, in the sense that the probability for DRS and FBS iterations with random initializations to converge to strict saddle points of their respective objectives is zero (see Theorem 6.5.6).

The main technical tools to achieve this are (i) Forward-Backward Envelope (FBE) [215], Douglas-Rachford Envelope (DRE) [171] from nonconvex analysis, and (ii) Stable-Center Manifold Theorem [206] from dynamical systems.

FBE and DRE are functions with nice properties even in the nonconvex settings. In Section 6.4, we show that they share the same stationary points, global minimizers, local minimizers, and strict saddle points with the objectives of FBS and DRS, respectively. Furthermore, the FBS and DRS iterations can be written as (preconditioned) gradient descent iterations on FBE and DRE. In Section 6.5, we analyze these gradient descent iterations with the Stable-Center Manifold Theorem, and show that whenever FBS and DRS converge, their limits will not be the strict saddles of FBE and DRE almost surely, which are exactly the strict saddles of their corresponding objective functions. Consequently, for many practical models that satisfy the *strict saddle property*[2], FBS and DRS will almost always avoid the strict saddle points whenever they converge.

As a byproduct, we also generalize FBE and DRE to the Davis-Yin Envelope in Section 6.3, which is an envelope function for the Davis-Yin splitting[3]. Many results in Section 6.4 and 6.5 also hold for Davis-Yin Splitting and Davis-Yin Envelope.

---

[1] I.e., saddle points with a negative curvature.

[2] That is, the stationary points of the objective are either local minimizers or strict saddle points.

[3] Davis-Yin splitting [69] is a generalization of FBS and DRS.

CHAPTER 6

Strict-Saddle Point Avoidance of FBS and DRS

6.1 Introduction

The most general model considered in this chapter minimizes the sum of three functions, where one of them is differentiable, and all the three functions can be nonconvex. A mathematical formulation is given in Section 6.3. The results of this chapter, of course, apply to simpler models, where any one or two of these three functions vanish. Problems that can be written in our general model are abundant. Examples include texture inpainting [134], matrix completion [48], and support vector machine classification [63].

Our model can be solved by the splitting iterative methods based on Douglas-Rachford Splitting (DRS) [133] and Forward-Backward Splitting (FBS) [168], as well as their generalization, Davis-Yin Splitting (DYS) [69]. In these methods, the problem objective is split into different steps, one for each of the objective functions. Their implementations are typically straightforward. By exploiting additional sum and coordinate friendly structures, they give rise to parallel and distributed algorithms that are highly scalable. The details of these methods are reviewed in Section 6.3 below.

These splitting methods are traditionally analyzed under the assumption that the subdifferentials of the objective functions are maximally monotone. The subdifferentials of nonconvex functions are generally non-monotone. Therefore, the majority of the existing results apply only to convex objective functions.

Recently, FBS and DRS are found to numerically converge for certain nonconvex problems, for example, FBS for image restoration [209], dictionary learning, and ma-

trix decomposition [215], and DRS for nonconvex feasibility problems [125], matrix completion [6], and phase retrieval [54]. Theoretically, their iterates have been shown to converge to stationary points in some nonconvex settings [10, 125, 216, 108]. In particular, any bounded sequence produced by FBS converges to a stationary point when the objective satisfies the KL property [10]; By using the Douglas-Rachford Envelope (DRE), the authors of [125] show that DRS iterates converge to a stationary point when one of the two functions is Lipschitz differentiable, both of them are semi-algebraic and bounded below, and one of them is coercive; Later, the boundedness assumption is removed in [216]; In [124], similar convergence is established for Peaceman-Rachford Splitting; In [108], when one function is strongly convex, and the other is weakly convex, and their sum is strongly convex, DRS iterates are shown to be Fejer monotone with respect to the set of fixed points of DRS operator, thus convergent. Though unlikely, it is still possible that the limit of a convergent sequence is a saddle point instead of a local minimum (except when all stationary points are local minima, which is the case studied in [108]).

On the other hand, some first-order methods have been shown to avoid so-called strict saddle points, with probability one regarding random initialization [120, 119]. These results make skillful use of the Stable-Center Manifold Theorem [206]. So far, their results apply only to relatively simple methods such as Gradient Descent, Coordinate Descent, and Proximal Point methods. We give an affirmative answer (under smoothness assumptions) that splitting methods also have this property. This result also matches the practical observations made in [213].

This chapter makes the following contribution regarding the envelopes and saddle point avoidance of FBS and DRS iterations for nonconvex problems. We first generalize the existing Forward-Backward Envelope (FBE) and Douglas-Rachford Envelope (DRE) into a Davis-Yin Envelope (DYE) and establish relationships between the latter envelope and the original optimization objective. Then, under smoothness conditions, we show that the probability for DRS and FBS iterations with random initializations

to converge to strict saddle points of their respective DRE and FBE is zero. Finally, by the connection between the envelopes and the original objectives, we extend the above avoidance results to the strict saddle points of the original objectives. That is, when our problem has the strict saddle property, DRS and FBS with random initialization will almost surely converge to local minimizers. The strict saddle property is satisfied in several applications including, but not limited to, dictionary learning [213], simple neural networks [46], phase retrieval [212], tensor decomposition [97], and low-rank matrix factorization [33].

Recently, another generalization of FBE and DRE is proposed[101]. Some properties of the more general envelope are provided and some of them sharpen the corresponding results of FBE and DRE. Also, a new interpretation of FBS and DRS as majorization-minimization algorithms applied to their respective envelopes is given. Compared to [101], the envelope proposed in this chapter also applies to DYS, we interpret DYS as gradient descent of this envelope under a variable metric, and establish the strict saddle avoidance property of FBS and DRS.

The rest of this chapter is organized as follows. In Section 6.2, we introduce notation and review some useful results. In Section 6.3, we review DYS, and define the envelope for DYS. In Section 6.4, we rewrite DYS equivalently as a gradient descent of the envelope, and establish a strong relationship between the envelope and the objective. Then, in Section 6.5, we analyze the avoidance of strict saddle points of the objective. Finally, we conclude this chapter in Section 6.6.

6.2 Preliminaries

In this section, we review some basic concepts, introduce our notation, and state some known results. For the sake of brevity, we omit proofs and direct references. We refer the reader to textbooks [194, 16].

We let $\mathbf{0} \in \mathbb{R}^n$ denote the vector zero, $\langle \cdot, \cdot \rangle$ the usual dot product, $\| \cdot \|$ the ℓ_2 norm, and Fix$T$ the set of fixed points of a single-valued operator T.

A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is called $\beta-$weakly convex (or $\beta-$semiconvex) if the function $\tilde{f}(\cdot) := f(\cdot) + \frac{\beta}{2}\| \cdot \|^2$ is convex. Clearly, f can be nonconvex.

Let $y \xrightarrow{f} x$ denote $y \to x$ and $f(y) \to f(x)$. Then, the subdifferential of f at $x \in \operatorname{dom} f$ can be defined by

$$\partial f(x) := \Big\{ v \in \mathbb{R}^n \,:\, \exists x^t \xrightarrow{f} x, v^t \to v, \text{with} \\ \liminf_{z \to x^t} \frac{f(z) - f(x^t) - \langle v^t, z - x^t \rangle}{\|z - x^t\|} \geq 0 \text{ for each } t \Big\}.$$

If f is differentiable at x, we have $\partial f(x) = \{\nabla f(x)\}$; If f is convex, we have

$$\partial f(x) = \{v \in \mathbb{R}^n \,:\, f(z) \geq f(x) + \langle v, z - x \rangle \text{ for any } z \in \mathbb{R}^n\},$$

which is the classic definition of subdifferential in convex analysis.

A point x^* is a stationary point of a function f if $\mathbf{0} \in \partial f(x^*)$. x^* is a critical point of f if f is differentiable at x^* and $\nabla f(x^*) = \mathbf{0}$.

A point x^* is a *strict saddle point* of f if f is twice differentiable at x^*, x^* is a critical point of f, and $\lambda_{\min}[\nabla^2 f(x^*)] < 0$, where $\lambda_{\min}[\cdot]$ returns the smallest eigenvalue of the input. Local minimizers of a function are always its stationary points, but not strict saddle points.

For any $\gamma > 0$, the Moreau envelope of a function f is defined by

$$f^\gamma(x) := \inf_{y \in \mathbb{R}^n} \{f(y) + \frac{1}{2\gamma}\|y - x\|^2\}.$$

The proximal mapping of f is defined by

$$\operatorname{Prox}_{\gamma f}(x) : \; x \rightrightarrows \arg\min_{y \in \mathbb{R}^n} \{f(y) + \frac{1}{2\gamma}\|y - x\|^2\},$$

assuming that the arg min exists, here $\rightrightarrows$ denotes a possibly set-valued mapping. When f is convex, $\operatorname{Prox}_{\gamma f}$ is single-valued and equals $\operatorname{Prox}_{\gamma f}(x) = (\mathrm{Id} + \gamma \partial f)^{-1}$, where Id is

the identity map. For any proper, closed, convex function f, its Moreau Identity is

$$\mathrm{Id} = \mathrm{Prox}_{\gamma f} + \gamma \, \mathrm{Prox}_{\frac{f^*}{\gamma}} \circ \frac{\mathrm{Id}}{\gamma}, \tag{6.1}$$

where $f^*(u) \coloneqq \sup_{x \in \mathbb{R}^n} \{\langle u, x \rangle - f(x)\}$ is the convex conjugate of f.

We also need the Inverse Function Theorem: let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a C^1 mapping, if the Jacobian $J_F(x)$ of F at $x \in \mathbb{R}^n$ is invertible, then, there exists an inverse function F^{-1} defined in a neighbourhood of $F(x)$ such that F^{-1} is also C^1 and

$$J_{F^{-1}}\Big(F(x)\Big) = \Big(J_F(x)\Big)^{-1}. \tag{6.2}$$

6.3 Davis-Yin Splitting and its Envelope

In this section, we will introduce a function, which we call an envelope, such that DYS iteration can be written as the gradient descent of this function under a variable metric. Since DYS generalizes FBS and DRS, the envelope of DYS is also a generalization of FBE and DRE, the respective envelopes of FBS and DRS, which were introduced in [171, 215].

6.3.1 Review of Davis-Yin Splitting

DYS [69] can be applied to solve the following problem:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \, \varphi(x) \coloneqq f(x) + g(x) + h(x), \tag{6.3}$$

where $f, g, h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$.

DYS iteration produces a sequence $(z^k)_{k \geq 0}$ according to $z^{k+1} = Tz^k$, where

$$Tz^k := z^k + \alpha\left(\mathrm{Prox}_{\gamma f}\left(2\,\mathrm{Prox}_{\gamma g}(z^k) - z^k - \gamma \nabla h\Big(\mathrm{Prox}_{\gamma g}(z^k)\Big)\right) - \mathrm{Prox}_{\gamma g}(z^k)\right),$$

where γ and α are positive scalars. We rewrite this operator into successive steps with

designated letters as

$$
\begin{aligned}
q^k &\coloneqq \nabla h\Big(\operatorname{Prox}_{\gamma g}(z^k)\Big),\\
r^k &\coloneqq 2\operatorname{Prox}_{\gamma g}(z^k) - z^k,\\
p^k &\coloneqq \operatorname{Prox}_{\gamma f}(r^k - \gamma q^k),\\
w^k &= p^k - \operatorname{Prox}_{\gamma g}(z^k), \qquad (6.4)\\
z^{k+1} &= Tz^k = z^k + \alpha w^k. \qquad (6.5)
\end{aligned}
$$

In [69], convergence is established when f, g and h are proper, closed, and convex, h is L_h-Lipschitz differentiable, and

$$\gamma \in]0, 2L_h[, \quad \alpha \in \Big]0, 2 - \frac{\gamma}{2L_h}\Big[.$$

When $h = 0$, (6.5) simplifies to DRS iteration,

$$z^{k+1} = z^k + \alpha\Big(\operatorname{Prox}_{\gamma f}(r^k) - \operatorname{Prox}_{\gamma g}(z^k)\Big).$$

When $g = 0$, $\operatorname{Prox}_{\gamma g}$ reduces to Id and thus (6.5) simplifies to

$$z^{k+1} = z^k + \alpha(\operatorname{Prox}_{\gamma f}(z^k - \gamma q^k) - z^k),$$

which is FBS iteration.

When $f = 0$, $\operatorname{Prox}_{\gamma f}$ reduces to Id and (6.5) simplifies to Backward-Forward Splitting,

$$z^{k+1} = z^k + \alpha(\operatorname{Prox}_{\gamma g}(z^k) - \gamma q^k - z^k).$$

When $f = g = 0$, (6.5) simplifies to gradient descent iteration

$$z^{k+1} = z^k - \alpha\gamma q^k.$$

6.3.2 Envelope of Davis-Yin Splitting

We define the envelope function of DYS as:

$$\varphi^\gamma(z) := g^\gamma(z) - \gamma\|\nabla g^\gamma(z)\|^2 - \gamma\langle \nabla h\Big(\operatorname{Prox}_{\gamma g}(z)\Big), \nabla g^\gamma(z)\rangle + h\Big(\operatorname{Prox}_{\gamma g}(z)\Big) - \frac{\gamma}{2}\|\nabla h\Big(\operatorname{Prox}_{\gamma g}(z)\Big)\|^2 + f^\gamma\Big(z - 2\gamma\nabla g^\gamma(z) - \gamma\nabla h\Big(\operatorname{Prox}_{\gamma g}(z)\Big)\Big). \tag{6.6}$$

When $g = 0$, this envelope reduces to the FBE proposed in [215]; When $h = 0$, it reduces to the DRE introduced in [171]. When $g = h = 0$, it is the Moreau envelope.

This envelope is well defined when g is $\beta-$weakly convex, h is differentiable, and $\gamma \in (0, \frac{1}{\beta})$. This is justified by the following lemma.

Lemma 6.3.1. *Let ξ be proper, closed, $\beta-$weakly convex. Choose γ such that $\gamma \in (0, \frac{1}{\beta})$. Let $\xi^\gamma(z) = \min_{u\in\mathbb{R}^n}\{\xi(u) + \frac{1}{2\gamma}\|z-u\|^2\}$ be the Moreau envelope of ξ. Define $\tilde{\xi}(\cdot) = \xi(\cdot) + \frac{\beta}{2}\|\cdot\|^2$, which is convex. Then, proximal mapping $\operatorname{Prox}_{\gamma\xi}(z)$ is single-valued and satisfies*

$$\operatorname{Prox}_{\gamma\xi}(z) = \operatorname{Prox}_{\frac{\gamma}{1-\gamma\beta}\tilde{\xi}}(\frac{1}{1-\gamma\beta}z),$$
$$\nabla\xi^\gamma(z) = \gamma^{-1}\Big(z - \operatorname{Prox}_{\gamma\xi}(z)\Big).$$

Furthermore, $\operatorname{Prox}_{\gamma\xi}(z)$ is $\frac{1}{1-\gamma\beta}-$Lipschitz continuous.

Proof. We have

$$\begin{aligned}\xi^\gamma(z) &= \min_{u\in\mathbb{R}}\{\xi(u) + \frac{\beta}{2}\|u\|^2 + \frac{1}{2\gamma}\|u-z\|^2 - \frac{\beta}{2}\|u\|^2\}\\ &= \min_{u\in\mathbb{R}}\Big\{\tilde{\xi}(u) + \frac{1-\gamma\beta}{2\gamma}\|u - \frac{1}{1-\gamma\beta}z\|^2\Big\} - \frac{\beta}{2-2\gamma\beta}\|z\|^2\end{aligned}$$

where the second equality follows from the definition of $\tilde{\xi}$.

As a result, for $\gamma \in (0, \frac{1}{\beta})$, $\operatorname{Prox}_{\gamma\xi}$ is single-valued and

$$\operatorname{Prox}_{\gamma\xi}(z) = \operatorname{Prox}_{\frac{\gamma}{1-\gamma\beta}\tilde{\xi}}(\frac{1}{1-\gamma\beta}z).$$

Since $\text{Prox}_{\frac{\gamma}{1-\gamma\beta}\tilde{\xi}}(z)$ is $1-$Lipschitz continuous, we know that $\text{Prox}_{\gamma\xi}(z)$ is Lipschitz continuous with constant $\frac{1}{1-\gamma\beta}$.

Finally, since $\tilde{\xi}$ is convex, [16, Prop.12.29] tells us that ξ^γ is differentiable and

$$\begin{aligned}\nabla\xi^\gamma(z) &= \frac{1}{1-\gamma\beta}\nabla\tilde{\xi}^{\frac{\gamma}{1-\gamma\beta}}\big(\frac{1}{1-\gamma\beta}z\big) - \frac{\beta}{1-\gamma\beta}z \\ &= \frac{1}{1-\gamma\beta}\frac{1-\gamma\beta}{\gamma}\Big(\frac{1}{1-\gamma\beta}z - \text{Prox}_{\frac{\gamma}{1-\gamma\beta}\tilde{\xi}}\big(\frac{1}{1-\gamma\beta}z\big)\Big) - \frac{\beta}{1-\gamma\beta}z \\ &= \frac{1}{\gamma}\Big(z - \text{Prox}_{\gamma\xi}(z)\Big).\end{aligned}$$

□

□

6.4 Properties of Envelope

In this section, we show that DYS iteration can be written as the gradient descent of this function under a variable metric. Furthermore, the global minimizers, local minimizers, critical (stationary) points, and strict saddle points of the envelope φ^γ defined in (6.6) correspond *one on one* to those of the objective function φ in (6.3).

We now analyze the properties of the DYS envelope (6.6) under the following assumption:

Assumption 6.4.1.

1. $g : \mathbb{R}^n \to \mathbb{R}$ *is* L_g-*Lipschitz differentiable.*
2. $h : \mathbb{R}^n \to \mathbb{R}$ *is* L_h-*Lipschitz differentiable.*
3. $\varphi : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ *is lower bounded.*
4. $\gamma \in (0, \frac{1}{L_g+L_h})$.

Compared to the assumption in Section 6.3.1, a main restriction is that g is Lipschitz differentiable. On the other hand, all f, g and h can be nonconvex.

First, we show lower and upper bounds of the envelope, which generalize [215, Prop. 2.3], [208, Prop. 4.3], and [171, Prop. 1].

Lemma 6.4.1. *Under Assumption 6.4.1, the following three inequalities hold for any* $z \in \mathbb{R}^n$*:*

$$\varphi^\gamma(z) \le \varphi\Big(\mathrm{Prox}_{\gamma g}(z)\Big), \tag{6.7}$$

$$\varphi^\gamma(z) \ge \varphi\Big(p(z)\Big) + C_1(\gamma)\|p(z) - \mathrm{Prox}_{\gamma g}(z)\|^2, \tag{6.8}$$

$$\varphi^\gamma(z) \le \varphi\Big(p(z)\Big) + C_2(\gamma)\|p(z) - \mathrm{Prox}_{\gamma g}(z)\|^2, \tag{6.9}$$

where $\varphi^\gamma(z)$ *is defined in* (6.6)*,*

$$C_1(\gamma) := \frac{1 - \gamma L_h - \gamma L_g}{2\gamma} > 0,$$

$$C_2(\gamma) := \frac{1 + \gamma L_h + \gamma L_g}{2\gamma} > 0,$$

and $p(z)$ *is any element of* $\mathrm{Prox}_{\gamma f}\Big(2\,\mathrm{Prox}_{\gamma g}(z) - z - \gamma\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big)\Big)$.

Proof of inequality (6.7). By applying Lemma 6.3.1 to g, $\varphi^\gamma(z)$ can be written as

$$\begin{aligned}
\varphi^\gamma(z) =& \min_u\{g(u) + \frac{1}{2\gamma}\|z - u\|^2\} - \gamma\|\frac{1}{\gamma}\Big(z - \mathrm{Prox}_{\gamma g}(z)\Big)\|^2 \\
&- \gamma\langle\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big), \frac{1}{\gamma}\Big(z - \mathrm{Prox}_{\gamma g}(z)\Big)\rangle \\
&+ h\Big(\mathrm{Prox}_{\gamma g}(z)\Big) - \frac{\gamma}{2}\|\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big)\|^2 \\
&+ \min_u\{f(u) + \frac{1}{2\gamma}\| - z + 2\,\mathrm{Prox}_{\gamma g}(z) - \gamma\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big) - u\|^2\}.
\end{aligned} \tag{6.10}$$

Taking $u = \mathrm{Prox}_{\gamma g}(z)$ in the two minimums of (6.10), we have

$$\begin{aligned}
\varphi^\gamma(z) \le& g\Big(\mathrm{Prox}_{\gamma g}(z)\Big) + \frac{1}{2\gamma}\|z - \mathrm{Prox}_{\gamma g}(z)\|^2 - \gamma\|\frac{1}{\gamma}\Big(z - \mathrm{Prox}_{\gamma g}(z)\Big)\|^2 \\
&- \langle\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big), z - \mathrm{Prox}_{\gamma g}(z)\rangle \\
&+ h\Big(\mathrm{Prox}_{\gamma g}(z)\Big) - \frac{\gamma}{2}\|\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big)\|^2 \\
&+ f\Big(\mathrm{Prox}_{\gamma g}(z)\Big) + \frac{1}{2\gamma}\| - z + \mathrm{Prox}_{\gamma g}(z) - \gamma\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big)\|^2 \\
=& \varphi\Big(\mathrm{Prox}_{\gamma g}(z)\Big).
\end{aligned}$$

□

□

Proof of inequality (6.8). According to Assumption 6.4.1, we know that $f(\cdot) + \frac{1}{2\gamma}\|\cdot\|^2$ is bounded below for $\gamma \in (0, \frac{1}{L_g+L_h})$. Therefore, [194, Thm. 1.25] tells us that $\mathrm{Prox}_{\gamma f}\Big(2\,\mathrm{Prox}_{\gamma g}(z) - z - \gamma\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big)\Big) \neq \varnothing$ for $\gamma \in (0, \frac{1}{L_g+L_h})$).

By taking $u = \mathrm{Prox}_{\gamma g}(z)$ in the first minimum of (6.10) and $u = p(z) \in \mathrm{Prox}_{\gamma f}(2\,\mathrm{Prox}_{\gamma g}(z) - z - \gamma\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big))$ in the second, we have

$$\begin{aligned}\varphi^\gamma(z) = {} & g\Big(\mathrm{Prox}_{\gamma g}(z)\Big) + \frac{1}{2\gamma}\|z - \mathrm{Prox}_{\gamma g}(z)\|^2 \\ & - \gamma\langle\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big), \frac{1}{\gamma}\Big(z - \mathrm{Prox}_{\gamma g}(z)\Big)\rangle \\ & + h\Big(\mathrm{Prox}_{\gamma g}(z)\Big) - \frac{\gamma}{2}\|\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big)\|^2 \\ & + f\Big(p(z)\Big) + \frac{1}{2\gamma}\| - z + 2\,\mathrm{Prox}_{\gamma g}(z) - \gamma\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big) - p(z)\|^2. \qquad (6.11)\end{aligned}$$

By making use of

$$h(y) \geq h(x) - \langle\nabla h(y), x - y\rangle - \frac{L_h}{2}\|x - y\|^2 \ \text{ forany } x, y \in \mathbb{R}^{\mathrm{n}},$$

we arrive at

$$\begin{aligned}\varphi^\gamma(z) \geq {} & g\Big(\mathrm{Prox}_{\gamma g}(z)\Big) - \frac{1}{2\gamma}\|z - \mathrm{Prox}_{\gamma g}(z)\|^2 \\ & - \langle\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big), z - \mathrm{Prox}_{P\gamma g}(z)\rangle \\ & + h\Big(p(z)\Big) - \langle\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big), (p(z) - \mathrm{Prox}_{\gamma g}(z))\rangle \\ & - \frac{L_h}{2}\|p(z) - \mathrm{Prox}_{\gamma g}(z)\|^2 - \frac{\gamma}{2}\|\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big)\|^2 \\ & + f\Big(p(z)\Big) + \frac{1}{2\gamma}\|2\,\mathrm{Prox}_{\gamma g}(z) - z - \gamma\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big) - p(z)\|^2.\end{aligned}$$

Next, by making use of $\|a+b+c\|^2 = \|a\|^2 + \|b\|^2 + \|c\|^2 + 2\langle a, b\rangle + 2\langle b, c\rangle + 2\langle a, c\rangle$ for

$$\begin{aligned}a &= \mathrm{Prox}_{\gamma g}(z) - p(z), \\ b &= \mathrm{Prox}_{\gamma g}(z) - z, \\ c &= -\gamma\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big),\end{aligned}$$

we obtain

$$\varphi^\gamma(z) \geq g\Big(\operatorname{Prox}_{\gamma g}(z)\Big) + h\Big(p(z)\Big) - \frac{L_h}{2}\|p(z) - \operatorname{Prox}_{\gamma g}(z)\|^2 \\ + f\Big(p(z)\Big) + \frac{1}{2\gamma}\|\operatorname{Prox}_{\gamma g}(z) - p(z)\|^2 + \langle \operatorname{Prox}_{\gamma g}(z) - p(z), \frac{1}{\gamma}(\operatorname{Prox}_{\gamma g}(z) - z)\rangle.$$

Finally, by substituting

$$\nabla g\Big(\operatorname{Prox}_{\gamma g}(z)\Big) = -\frac{1}{\gamma}(\operatorname{Prox}_{\gamma g}(z) - z),$$

$$g(y) \geq g(x) - \langle \nabla g(y), x - y\rangle - \frac{L_g}{2}\|x - y\|^2 \quad \text{forany } x, y \in \mathbb{R}^n,$$

we arrive at (6.8). □ □

Proof of inequality (6.9). Similarly to the proof above, we can also start from (6.11) and apply

$$h(y) \leq h(x) - \langle \nabla h(y), x - y\rangle + \frac{L_g}{2}\|x - y\|^2 \text{ forany } x, y \in \mathbb{R}^n,$$

$$g(y) \leq g(x) - \langle \nabla g(y), x - y\rangle + \frac{L_g}{2}\|x - y\|^2 \text{ forany } x, y \in \mathbb{R}^n,$$

which gives (6.9). □ □

6.4.1 Global Minimizers Correspondence

Now we can establish the direct connections between the global and local minimizers of φ^γ and those of φ. These results generalize [215, Prop. 2.3] and [208, Thm. 4.4].

Theorem 6.4.2. *Under Assumption 6.4.1, we have*

1. $\inf_{x\in\mathbb{R}^n} \varphi(x) = \inf_{z\in\mathbb{R}^n} \varphi^\gamma(z)$,
2. $\arg\min_{x\in\mathbb{R}^n} \varphi(x) = \operatorname{Prox}_{\gamma g}\Big(\arg\min_{z\in\mathbb{R}^n}\Big(\varphi^\gamma(z)\Big)\Big)$.

Proof of 1. From (6.7) we have

$$\inf_{z\in\mathbb{R}^n} \varphi^\gamma(z) \leq \inf_{x\in\mathbb{R}^n} \varphi(x),$$

If $\inf_{z\in\mathbb{R}^n}\varphi^\gamma(z) < \inf_{x\in\mathbb{R}^n}\varphi(x)$, then, there exists $z_1\in\mathbb{R}^n$ such that $\varphi^\gamma(z_1) < \inf_{x\in\mathbb{R}^n}\varphi(x)$. So (6.8) gives

$$\inf_{x\in\mathbb{R}^n}\varphi(x) > \varphi^\gamma(z_1) \geq \varphi\Big(p(z_1)\Big) + C_1(\gamma)\|\operatorname{Prox}_{\gamma g}(z_1) - p(z_1)\|^2 \geq \varphi\Big(p(z_1)\Big),$$

which is a contradiction. □ □

Proof of 2. Let us first show that

$$\arg\min_{x\in\mathbb{R}^n}\varphi(x) \subseteq \operatorname{Prox}_{\gamma g}\Big(\arg\min_{z\in\mathbb{R}^n}\Big(\varphi^\gamma(z)\Big)\Big).$$

Without the loss of generality, we may assume $\arg\min_{x\in\mathbb{R}^n}\varphi(x)\neq\varnothing$, then, for any $x^*\in\arg\min_{x\in\mathbb{R}^n}\varphi(x)$, we have $x^* = \operatorname{Prox}_{\gamma g}(z^*)$ for $z^* = (I+\gamma\nabla g)(x^*)$. As a result, (6.7) and (6.8) give us

$$\begin{aligned}\inf_{x\in\mathbb{R}^n}\varphi(x) = \varphi(x^*) &= \varphi\Big(\operatorname{Prox}_{\gamma g}(z^*)\Big)\\ &\geq \varphi^\gamma(z^*) \geq \varphi\Big(p(z^*)\Big) + C_1(\gamma)\|\operatorname{Prox}_{\gamma g}(z^*) - p(z^*)\|^2,\end{aligned}$$

which enforces $\operatorname{Prox}_{\gamma g}(z^*) = p(z^*)$ and $\varphi\Big(\operatorname{Prox}_{\gamma g}(z^*)\Big) = \varphi^\gamma(z^*)$. So for any $z\in\mathbb{R}^n$ we have

$$\varphi^\gamma(z^*) = \inf_{x\in\mathbb{R}^n}\varphi(x) \leq \varphi\Big(p(z)\Big) \leq \varphi^\gamma(z) - C_1(\gamma)\|\operatorname{Prox}_{\gamma g}(z) - p(z)\|^2 \leq \varphi^\gamma(z),$$

which yields $z^*\in\arg\min_{z\in\mathbb{R}^n}\varphi^\gamma(z), x^*\in\operatorname{Prox}_{\gamma g}\Big(\arg\min_{z\in\mathbb{R}^n}\Big(\varphi^\gamma(z)\Big)\Big)$.

Now let us show that

$$\operatorname{Prox}_{\gamma g}\Big(\arg\min_{z\in\mathbb{R}^n}\Big(\varphi^\gamma(z)\Big)\Big) \subseteq \arg\min_{x\in\mathbb{R}^n}\varphi(x).$$

Again, we can assume that $\arg\min_{z\in\mathbb{R}^n}\Big(\varphi^\gamma(z)\Big)\neq\varnothing$. For any $z^*\in\arg\min_{z\in\mathbb{R}^n}\varphi^\gamma(z)$, we need to show $\operatorname{Prox}_{\gamma g}(z^*)\in\arg\min_{x\in\mathbb{R}^n}\varphi(x)$.

Let $z^{**} = (I+\gamma\nabla g)p(z^*)$, then, $\operatorname{Prox}_{\gamma g}(z^{**}) = p(z^*)$ and (6.7) and (6.8) give us

$$\varphi^\gamma(z^{**}) \leq \varphi(\operatorname{Prox}_{\gamma g}\Big(z^{**}\Big)\Big) = \varphi\Big(p(z^*)\Big) \leq \varphi^\gamma(z^*) - C_1(\gamma)\|\operatorname{Prox}_{\gamma g}(z^*) - p(z^*)\|^2.$$

Since $z^* \in \arg\min_{z\in\mathbb{R}^n} \varphi^\gamma(z)$, we must have

$$\mathrm{Prox}_{\gamma g}(z^*) = p(z^*) = \mathrm{Prox}_{\gamma g}(z^{**}),$$

$$\varphi^\gamma(z^*) = \varphi^\gamma(z^{**}) = \varphi\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big).$$

Consequently, for any $z \in \mathbb{R}^n$ we have

$$\varphi\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big) = \varphi^\gamma(z^*) \le \varphi^\gamma(z) \le \varphi\Big(\mathrm{Prox}_{\gamma g}(z)\Big),$$

which concludes $\mathrm{Prox}_{\gamma g}(z^*) \in \arg\min_{x\in\mathbb{R}^n} \varphi(x)$. □ □

6.4.2 Davis-Yin Splitting as Gradient Descent of the Envelope

We now show that (6.5) can be written as a gradient descent iteration of an envelope function under the following assumption.

Assumption 6.4.2.

1. *f is β_f-weakly convex and $\gamma \in (0, \frac{1}{\beta_f})$.*
2. *g, h are twice continuously differentiable.*

We begin with a technical lemma regarding the twice differentiability of the Moreau envelope of g.

Lemma 6.4.3. *Under Assumptions 6.4.1 and 6.4.2, $\mathrm{Prox}_{\gamma g}$ has a Jacobian at z^0, g^γ is twice differentiable at z^0 with the Hessian*

$$\nabla^2 g^\gamma(z^0) = \frac{1}{\gamma}\left(I - \Big(I + \gamma\nabla^2 g\Big(\mathrm{Prox}_{\gamma g}(z^0)\Big)\Big)^{-1}\right).$$

In addition, the mapping

$$A(z) \coloneqq I - 2\gamma\nabla^2 g^\gamma(z) - \gamma\nabla^2 h\Big((\mathrm{Prox}_{\gamma g}(z)\Big)\Big(I - \gamma\nabla^2 g^\gamma(z)\Big) \tag{6.12}$$

is invertible.

Proof. Since $\gamma \in]0, \frac{1}{L_g}[$, $\text{Prox}_{\gamma g}$ is single-valued and

$$\text{Prox}_{\gamma g}(z^0) = (\text{Id} + \gamma \nabla g)^{-1} z^0, \tag{6.13}$$

where $(\text{Id} + \gamma \nabla g)^{-1}$ is the inverse mapping of $\text{Id} + \gamma \nabla g$. Since $\nabla^2 g(\text{Prox}_{\gamma g}\left(z^0\right))$ is symmetric and its eigenvalues are bounded by L_g, we know that
$I + \gamma \nabla^2 g\Big(\text{Prox}_{\gamma g}(z^0)\Big)$ is invertible, which is the Jacobian of $\text{Id} + \gamma \nabla g$ at $\text{Prox}_{\gamma g}(z^0)$.

Applying the Inverse Function Theorem to (6.13) by setting F as $\text{Prox}_{\gamma g}$ and z^0 as p in (6.2), we have

$$J_{\text{Prox}_{\gamma g}}(z^0) = \Big(I + \gamma \nabla^2 g\Big(\text{Prox}_{\gamma g}(z^0)\Big)\Big)^{-1}$$

Hence, Lemma 6.3.1 yields

$$\nabla^2 g^\gamma(z^0) = \frac{1}{\gamma}\Big(I - \Big(I + \gamma \nabla^2 g\Big(\text{Prox}_{\gamma g}(z^0)\Big)\Big)^{-1}\Big).$$

According to (6.12),

$$A(z^0) = A_1 - \gamma A_2. \tag{6.14}$$

where

$$\begin{aligned} A_1 &= 2\Big(I + \gamma \nabla^2 g\Big(\text{Prox}_{\gamma g}(z^0)\Big)\Big)^{-1} - I, \\ A_2 &= \nabla^2 h\Big(\text{Prox}_{\gamma g}(z^0)\Big)\Big(I + \gamma \nabla^2 g\Big(\text{Prox}_{\gamma g}(z^0)\Big)\Big)^{-1}. \end{aligned}$$

Since $\gamma \in (0, \frac{1}{L_g})$, A_1 is invertible, as a result,

$$\begin{aligned} \det\Big(A(z^0)\Big) &= \det(A_1 - \gamma A_2) = \det(I - \gamma A_2 A_1^{-1})\det(A_1) \\ &= \prod_{i=1}^{n}(1 - \gamma \lambda_i\Big(A_2 A_1^{-1}\Big)\Big)\det(A_1), \end{aligned}$$

where $\lambda_i(A_2 A_1^{-1}), i = 1, ..., n$ are the eigenvalues of $A_2 A_1^{-1}$.

Let us set

$$C = I + \gamma \nabla^2 g\Big(\text{Prox}_{\gamma g}(z^0)\Big) = C^T \succ 0,$$

and rewrite $A_2A_1^{-1}$ as

$$A_2A_1^{-1} = \nabla^2 h\Big(\operatorname{Prox}_{\gamma g}(z^0)\Big)C^{-1}(2C^{-1}-I)^{-1} = \nabla^2 h\Big(\operatorname{Prox}_{\gamma g}(z^0)\Big)(2I-C)^{-1}.$$

Note that $\nabla^2 h\Big(\operatorname{Prox}_{\gamma g}(z^0)\Big)$ is symmetric and $(2I-C)^{-1}$ is symmetric, positive definite. Therefore, $\lambda_i(A_2A_1^{-1}) \in \mathbb{R}$, and we can set

$$\lambda_1(A_2A_1^{-1}) \geq \lambda_2(A_2A_1^{-1}) \geq ... \geq \lambda_n(A_2A_1^{-1}).$$

In order to show $\det\Big(A(z^0)\Big) \neq 0$, it suffices to show that $1-\gamma\lambda_1(A_2A_1^{-1}) > 0$ when $\gamma \in (0, \frac{1}{L_g+L_h})$.

We have

$$\begin{aligned}
\lambda_1(A_2A_1^{-1}) &\overset{(a)}{\leq} \lambda_1\Big(\nabla^2 h\Big(\operatorname{Prox}_{\gamma g}(z^0)\Big)\Big)\cdot\lambda_1\Big((2I-C)^{-1}\Big) \\
&\leq \lambda_1\Big(\nabla^2 h\Big(\operatorname{Prox}_{\gamma g}(z^0)\Big)\Big)\cdot\frac{1}{2-(1+\gamma L_g)} \\
&= \|(\nabla^2 h\Big(\operatorname{Prox}_{\gamma g}(z^0)\Big)\|_2\cdot\frac{1}{1-\gamma L_g} \\
&\overset{(b)}{\leq} \|\nabla^2 h\Big(\operatorname{Prox}_{\gamma g}(z^0)\Big)\|_2\frac{1}{1-\gamma L_g} \\
&\leq L_h\frac{1}{1-\gamma L_g},
\end{aligned}$$

where (a) is by [242, Corollary 11], and (b) is by Cauchy-Schwartz. Since $\gamma \in (0, \frac{1}{L_g+L_h})$, we have

$$1-\gamma\lambda_1(A_2A_1^{-1}) \geq 1 - L_h\frac{\gamma}{1-\gamma L_g} > 0.$$

Therefore, $\det\Big(A(z^0)\Big) \neq 0$. □ □

Theorem 6.4.4. *Under Assumptions 6.4.1 and 6.4.2, DYS iteration (6.5) can be written equivalently as*

$$z^{k+1} = T(z^k) = z^k - \alpha\gamma A^{-1}(z^k)\nabla\varphi^\gamma(z^k), \tag{6.15}$$

where the metric is given by

$$A(z) := I - 2\gamma\nabla^2 g^\gamma(z) - \gamma\nabla^2 h\Big(\operatorname{Prox}_{\gamma g}(z)\Big)\Big(I-\gamma\nabla^2 g^\gamma(z)\Big),$$

and the envelope φ^γ is given by (6.6).

Proof. In view of Lemma 6.3.1 and (6.4), we have

$$w^k = p(z^k) - \mathrm{Prox}_{\gamma g}(z^k) = \mathrm{Prox}_{\gamma f}(r^k - \gamma q^k) - \mathrm{Prox}_{\gamma g}(z^k), \tag{6.16}$$

where

$$\begin{aligned}
\mathrm{Prox}_{\gamma f}(r^k - \gamma q^k) &= r^k - \gamma q^k - \gamma \nabla f^\gamma(r^k - \gamma q^k). \\
\mathrm{Prox}_{\gamma g}(z^k) &= z^k - \gamma \nabla g^\gamma(z^k), \\
r^k &= 2\,\mathrm{Prox}_{\gamma g}(z^k) - z^k = z^k - 2\gamma \nabla g^\gamma(z^k), \\
q^k &= q(z^k) = \nabla h\Big(\big(z^k - \gamma \nabla g^\gamma(z^k)\big)\Big).
\end{aligned}$$

By substitution,

$$w^k = -\gamma \nabla g^\gamma(z^k) - \gamma q(z^k) - \gamma \nabla f^\gamma\Big(z^k - 2\gamma \nabla g^\gamma(z^k) - \gamma q(z^k)\Big).$$

Let ∇_z denote taking gradient to z; then,

$$\begin{aligned}
&\nabla_z f^\gamma\Big(z - 2\gamma \nabla g^\gamma(z) - \gamma \nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big)\Big) \\
&= A(z) \nabla f^\gamma\Big(z - 2\gamma \nabla g^\gamma(z) - \gamma \nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big)\Big),
\end{aligned}$$

where $A(z)$ is given in (6.12). After some computation, we can verify that

$$\begin{aligned}
A(z^k) w^k =& -\gamma\Big(\nabla_z g^\gamma(z^k) - \gamma \nabla_z \|\nabla g^\gamma(z^k)\|^2\Big) \\
& -\gamma\Big(-\gamma \nabla_z\Big(\langle \nabla h\Big(\mathrm{Prox}_{\gamma g}(z^k)\Big), \nabla g^\gamma(z^k)\rangle\Big)\Big) \\
& -\gamma \nabla_z h\Big(\mathrm{Prox}_{\gamma g}(z^k)\Big) - \gamma\Big(-\frac{\gamma}{2} \nabla_z \|\nabla h\Big(\mathrm{Prox}_{\gamma g}(z^k)\Big)\|^2\Big) \\
& --\gamma \nabla_z f^\gamma\Big(z^k - 2\gamma \nabla g^\gamma(z^k) - \gamma \nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big)\Big) \\
=& -\gamma \nabla \varphi^\gamma(z^k).
\end{aligned}$$

Since $A(z^k)$ is invertible, we can rewrite DYS iteration (6.5) as (6.15). □ □

6.4.3 Local Minimizers Correspondence

Theorem 6.4.5. *Under Assumptions 6.4.1 and 6.4.2, we have:*

1. *If* $\mathrm{Prox}_{\gamma g}(z^*) \in \arg\min_{x \in B(\mathrm{Prox}_{\gamma g}(z^*), \delta)} \varphi(x)$ *for some* $\delta > 0$*, then,* z^* *is a local minimizer of* φ^γ*.*

2. *If* $z^* \in \arg\min_{z \in B(z^*, \varepsilon)} \varphi^\gamma(z)$ *for some* $\varepsilon > 0$*, then,*

$$\varphi\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big) \leq \varphi\Big(\mathrm{Prox}_{\gamma g}(z)\Big) \text{ for all } z \text{ such that } \|z - z^*\| \leq \varepsilon.$$

That is, $\mathrm{Prox}_{\gamma g}(z^*)$ *is a local minimizer of* $\varphi(x)$*.*

Proof of 1. Since $\mathrm{Prox}_{\gamma g}(z^*)$ is a local minimizer of φ, according to [194, Exercise 10.10], we have

$$\mathbf{0} \in \partial\varphi\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big) = \partial f\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big) + \nabla g\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big) + \nabla h\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big).$$

Since $\mathrm{Prox}_{\gamma g}$ is single-valued, this is equivalent to

$$\mathbf{0} \in \partial f\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big) + \frac{1}{\gamma}\Big(- \mathrm{Prox}_{\gamma g}(z^*) + z^* + \gamma \nabla h\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big)\Big),$$

Since $f + \frac{1}{2\gamma}\|\cdot\|^2$ is convex and $\mathrm{Prox}_{\gamma f}$ is single valued, this is further equivalent to

$$\mathrm{Prox}_{\gamma g}(z^*) = \mathrm{Prox}_{\gamma f}\left(2\,\mathrm{Prox}_{\gamma g}(z^*) - z^* - \gamma \nabla h\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big)\right) = p(z^*).$$

According to Lemma 6.3.1, $\mathrm{Prox}_{\gamma f}$ is $\frac{1}{1-\gamma\beta_f}-$Lipschitz continuous, we can conclude that there exists $\eta > 0$ such that when $\|z - z^*\| \leq \eta$, we have $\|p(z) - p(z^*)\| \leq \delta$ and

$$\begin{aligned}
\varphi^\gamma(z^*) = \varphi\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big) = \varphi\Big(p(z^*)\Big) &\leq \varphi\Big(p(z)\Big) \\
&\leq \varphi^\gamma(z) - C_1(\gamma)\|\,\mathrm{Prox}_{\gamma g}(z) - p(z)\|^2 \\
&\leq \varphi^\gamma(z).
\end{aligned}$$

□

□

Proof of 2. According to Lemma 6.4.3, $A(z)$ is invertible at z^*. Theorem 6.4.4 tells us that φ^γ is differentiable at z^*, so $\nabla\varphi^\gamma(z^*) = \mathbf{0}$ and $\mathrm{Prox}_{\gamma g}(z^*) = p(z^*)$. As a result, for any $z \in \mathbb{R}^n$ with $\|z - z^*\| \le \varepsilon$ we have

$$\varphi\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big) = \varphi^\gamma(z^*) \le \varphi^\gamma(z) \le \varphi\Big(\mathrm{Prox}_{\gamma g}(z)\Big).$$

Furthermore, according to Lemma 6.4.3 we have

$$\mathrm{Prox}_{\gamma g}(z) = \mathrm{Prox}_{\gamma g}(z^*) + \Big(I + \gamma\nabla^2 g\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big)\Big)^{-1}(z - z^*) + o(\|z - z^*\|).$$

Since $\Big(I + \gamma\nabla^2 g\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big)\Big)^{-1}$ is positive definite, we know that $\mathrm{Prox}_{\gamma g}\Big(B(z^*, \varepsilon)\Big)$ contains a ball centered at $\mathrm{Prox}_{\gamma g}(z^*)$, as a result, $\mathrm{Prox}_{\gamma g}(z^*)$ is a local minimizer of $\varphi(x)$. □ □

Now, let us show the one-to-one correspondence between the critical points of the envelope φ^γ and the stationary points of the objective $\varphi(x)$.

6.4.4 Critical and Stationary Point Correspondence

Theorem 6.4.6. *Under Assumptions 6.4.1 and 6.4.2, z^* is a critical point of φ^γ if and only if $\mathrm{Prox}_{\gamma g}(z^*)$ is a stationary point of φ.*

Proof. Since f is β_f-weakly convex and $\gamma \in (0, \frac{1}{\beta_f})$, by Lemma 6.3.1, we know that $\mathrm{Prox}_{\gamma f}$ is single-valued. And by Theorem 6.4.4, we have

$$\nabla\varphi^\gamma(z) = -A(z)\frac{1}{\gamma}(p(z) - \mathrm{Prox}_{\gamma g}(z)), \tag{6.17}$$

where $p(z) = \mathrm{Prox}_{\gamma f}\Big((2\,\mathrm{Prox}_{\gamma g}(z) - z - \gamma\nabla h\Big(\mathrm{Prox}_{\gamma g}(z)\Big)\Big)$. So $\nabla\varphi^\gamma(z^*) = \mathbf{0}$ if and only if

$$\begin{aligned}\mathrm{Prox}_{\gamma g}(z^*) &= \mathrm{Prox}_{\gamma f}\Big(2\,\mathrm{Prox}_{\gamma g}(z^*) - z^* - \gamma\nabla h(\mathrm{Prox}_{\gamma g}(z^*))\Big)\\ &= \arg\min_z\{f(z) + \frac{1}{2\gamma}\|z - \Big(2\,\mathrm{Prox}_{\gamma g}(z^*) - z^* - \gamma\nabla h\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big)\Big)\|^2\}.\end{aligned}$$

Since the objective in the arg min is convex, by [194, Exercise 10.10] we know that this is equivalent to

$$\mathbf{0} \in \partial f\Big(\operatorname{Prox}_{\gamma g}(z^*)\Big) + \frac{1}{\gamma}\Big(- \operatorname{Prox}_{\gamma g}(z^*) + z^* + \gamma \nabla h(\operatorname{Prox}_{\gamma g}(z^*))\Big).$$

By the definition of $\operatorname{Prox}_{\gamma g}$ and $\gamma \in (0, \frac{1}{L_g + L_h})$, this is further equivalent to

$$\mathbf{0} \in \partial f\Big(\operatorname{Prox}_{\gamma g}(z^*)\Big) + \nabla g\Big(\operatorname{Prox}_{\gamma g}(z^*)\Big) + \nabla h\Big(\operatorname{Prox}_{\gamma g}(z^*)\Big) = \partial \varphi\Big(\operatorname{Prox}_{\gamma g}(z^*)\Big). \qquad \square$$

$\square$

6.4.5 Strict Saddle Correspondence

In order to establish the correspondence between the strict saddles of φ^γ and φ, we also need the following assumption.

Assumption 6.4.3. *For any critical point z^* of φ^γ, f is twice continuously differentiable in a small neighbourhood of $\operatorname{Prox}_{\gamma g}(z^*)$, and there exits $L_f > 0$ such that $\|\nabla^2 f(\operatorname{Prox}_{\gamma g}(z^*))\| \le L_f$. In addition, assume that $\gamma \in (0, \frac{1}{L_f})$.*

Lemma 6.4.7. *Let z^* be a critical point of φ^γ. Under Assumptions 6.4.1, 6.4.2 and 6.4.3, φ^γ is twice differentiable at z^* and*

$$\nabla^2 \varphi^\gamma(z^*) = -A(z^*)\frac{1}{\gamma}\Big(J_p(z^*) - J_{\operatorname{Prox}_{\gamma g}}(z^*)\Big) \tag{6.18}$$

Moreover, $\nabla^2 \varphi^\gamma(z^*)$ is symmetric.

Proof. (6.18) follows from (6.17), $p(z^*) = \operatorname{Prox}_{\gamma g}(z^*)$, and [208, Prop. 2.A.2].

Since f is weakly convex, by Lemma 6.3.1 we know that $\operatorname{Prox}_{\gamma f}$ is continuous, so $p(z)$ is continuous. As a result, $\varphi^\gamma(z)$ is C^1, which tells us that $\nabla^2 \varphi^\gamma(z^*)$ is symmetric.

$\square$ $\square$

Theorem 6.4.8. *Let z^* be a critical point of φ^γ. Under Assumptions 6.4.1, 6.4.2 and 6.4.3, z^* is a strict saddle point of φ^γ if and only if $\operatorname{Prox}_{\gamma g}(z^*)$ is a strict saddle point of φ.*

Proof. According to Lemma 6.4.7, we know that $\nabla^2\varphi^\gamma(z^*)$ exists and it is symmetric.

Let z^* be a strict saddle point of $\varphi^\gamma(z)$, then, Taylor expansion gives

$$\begin{aligned}\varphi^\gamma(z) =&\varphi^\gamma(z^*) + \frac{1}{2}(z-z^*)^T\nabla^2\varphi^\gamma(z^*)(z-z^*) + o(\|z-z^*\|^2),\\ \varphi\Big(p(z)\Big) =&\varphi\Big(p(z^*)\Big) + \frac{1}{2}\Big(p(z)-p(z^*)\Big)^T\nabla^2\varphi\Big(p(z^*)\Big)\Big(p(z)-p(z^*)\Big)\\ &+ o(\|p(z)-p(z^*)\|^2).\end{aligned}$$

On the other hand, (6.8) gives

$$\varphi^\gamma(z) \geq \varphi\Big(p(z)\Big).$$

Let $\nabla^2\varphi^\gamma(z^*)v = \lambda v$, where $\|v\| = 1$ and $\lambda < 0$. Setting $z - z^* = \alpha v$, we arrive at

$$\begin{aligned}&\varphi^\gamma(z^*) + \frac{1}{2}\lambda\alpha^2 + o(\alpha^2)\\ &\geq \varphi\Big(p(z^*)\Big) + \frac{1}{2}\Big(p(z)-p(z^*)\Big)^T\nabla^2\varphi\Big(p(z^*)\Big)\Big(p(z)-p(z^*)\Big)\\ &\quad + o(\|p(z)-p(z^*)\|^2)\end{aligned} \tag{6.19}$$

Furthermore, (6.7), (6.8) together with $p(z^*) = \mathrm{Prox}_{\gamma g}(z^*)$ yield $\varphi^\gamma(z^*) = \varphi\Big(p(z^*)\Big)$, combine this with (6.19) and $\|p(z)-p(z^*)\| = O(\|z-z^*\|) = O(\alpha)$, we conclude that $\lambda_{\min}\Big(\nabla^2\varphi\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big)\Big) < 0$.

Similarly, let $\mathrm{Prox}_{\gamma g}(z^*)$ be a strict saddle of $\varphi(z)$, then, Taylor expansions gives

$$\begin{aligned}\varphi^\gamma(z) =&\varphi^\gamma(z^*) + \frac{1}{2}(z-z^*)^T\nabla^2\varphi^\gamma(z^*)(z-z^*) + o(\|z-z^*\|^2),\\ \varphi(\mathrm{Prox}_{\gamma g}\left(z\right)) =&\varphi\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big)\\ &+ \frac{1}{2}\Big(\mathrm{Prox}_{\gamma g}(z) - \mathrm{Prox}_{\gamma g}(z^*)\Big)^T\nabla^2\varphi\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big)\Big(\mathrm{Prox}_{\gamma g}(z) - \mathrm{Prox}_{\gamma g}(z^*)\Big)\\ &+ o(\|\mathrm{Prox}_{\gamma g}(z) - \mathrm{Prox}_{\gamma g}(z^*)\|^2).\end{aligned}$$

On the other hand, (6.7) gives

$$\varphi^\gamma(z) \leq \varphi\Big(\mathrm{Prox}_{\gamma g}(z)\Big),$$

Let $\nabla^2\varphi\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big)v = \lambda v$ where $\|v\| = 1$ and $\lambda < 0$ is a negative eigenvalue of $\nabla^2\varphi\Big(\mathrm{Prox}_{\gamma g}(z^*)\Big)$, let us also set $z = (\mathrm{Id} + \gamma\nabla g)(\mathrm{Prox}_{\gamma g}(z^*) + \alpha v)$. Therefore, $\mathrm{Prox}_{\gamma g}(z) - \mathrm{Prox}_{\gamma g}(z^*) = \alpha v$, taking $\alpha \to 0$ gives $\lambda_{\min}\Big(\nabla^2\varphi^\gamma(z^*)\Big) < 0$. □ □

6.5 Avoidance of Strict Saddle Points

In this section, we first show that under Assumptions 6.4.1, 6.4.2 and 6.4.3, the probability for DRS and FBS with random initializations to converge to strict saddle points of DRE and FBE is zero, respectively. Then, by combining this result with the correspondence between the strict saddle points of the envelope and the objective, as stated in Theorem 6.4.8, we can conclude that DRS and FBS, if convergent, will almost always avoid the strict saddle points of the objective. Therefore, when the objective satisfies the "strict saddle property", DRS and FBS, if they converge, will almost always converge to local minimizers.

To prove the main result, Theorem 6.5.6, we need the following Stable-Center Manifold Theorem, and its direct consequence, Theorem 6.5.2.

Theorem 6.5.1 states that, if T is a local diffeomorphism around one of its fixed point z^*, then, there is a local stable center manifold $W^{\text{cs}}_{\text{loc}}$ with dimension equal to the number of eigenvalues of the Jacobian of T at z^* that are less than or equal to 1. Furthermore, there exists a neighbourhood B of z^*, such that a point z must be in $W^{\text{cs}}_{\text{loc}}$ if its forward iterations $T^k(z)$, for all $k \geq 0$, stay in B.

Theorem 6.5.1 (Theorem III.7, Shub [206])**.** *Let z^* be a fixed point for a C^r local diffeomorphism $T : U \to \mathbb{R}^{\text{n}}$, where U is a neighbourhood of z^* and $r \geq 1$. Suppose $E = E_s \bigoplus E_u$, where E_s is the span of the eigenvectors that correspond to eigenvalues of $J_T(z^*)$ that have magnitude less than, or equal to, 1, and E_u is the span of eigenvectors that correspond to eigenvalues of $J_T(z^*)$ that have magnitude greater than 1. Then, there exists a C^r embedded disk $W^{\text{cs}}_{\text{loc}}$ that is tangent to E_s at z^*, which is called the local stable center manifold. Moreover, there exists a neighbourhood B of z^*, such that $T(W^{\text{cs}}_{\text{loc}}) \cap B \subseteq W^{\text{cs}}_{\text{loc}}$ and $\cap_{k=0}^{\infty} T^{-k}(B) \subseteq W^{\text{cs}}_{\text{loc}}$, where $T^{-k}(B) = \{z \in \mathbb{R}^{\text{n}} : T^k(z) \in B\}$.*

The assumption of this following theorem is weaker than that of Theorem 2 of [119], as we do not assume that T is invertible in $\mathbb{R}^{\text{n}}$ but only around every $z^* \in A^*_T$.

Theorem 6.5.2. *Assume that* $T(z) = z+\alpha\Big(p(z)-\mathrm{Prox}_{\gamma g}(z)\Big)$ *is a local diffeomorphism around every* $z^* \in A_T^* \coloneqq \{z \in \mathbb{R}^{\mathrm{n}} : z = T(z), \max_i |\lambda_i\Big(J_T(z)\Big)| > 1\}$, *where* A_T^* *is the set of unstable fixed points of* T, *and* $|\cdot|$ *denotes the magnitude. Then, the set* $W = \{z^0 : \lim z^k \in A_T^*\}$ *has Lebesgue measure* $\mu(W) = 0$ *in* $\mathbb{R}^{\mathrm{n}}$.

Proof. Take any $z^0 \in W$, we have $z^k = T^k(z^0) \to z^* \in A_T^*$, there exists $t_0 > 0$, such that for any $t \geq t_0$ we have $T^t(z^0) \in B_{z^*}$. As a result,
$T^t(z^0) \in S \coloneqq \cap_{k=0}^{\infty} T^{-k}(B_{z^*})$ for any $t \geq t_0$.

From Theorem 6.5.1 we know that S is a subset of the local center stable manifold $W_{\mathrm{loc}}^{\mathrm{cs}}$ whose codimension is greater or equal to 1, so we have $\mu(S) = 0$;

Finally, $T^{t_0}(z^0) \in S$ implies that $z^0 \in T^{-t_0}(S) \subseteq \cup_{j=0}^{\infty} T^{-j}(S)$, since

$$T^{-j}(S) = T^{-j}\Big(\cap_{k=0}^{\infty} T^{-k}(B_{z^*})\Big) = \cap_{k=j}^{\infty} T^{-k}(B_{z^*}) \subseteq \cap_{k=0}^{\infty} T^{-k}(B_{z^*}) = S,$$

we can conclude that $\mu(W) = 0$. □ □

Now let us show that $T(z)$ in Theorem 6.5.2 is indeed a local diffeomorphism around its fixed points.

Lemma 6.5.3. *Let* $T(z) = z+\alpha\Big(p(z)-\mathrm{Prox}_{\gamma g}(z)\Big)$ *and* $z^* \in \mathrm{Fix}T$. *Under Assumptions 6.4.1, 6.4.2 and 6.4.3, there exists* $\alpha_0 > 0$, *such that* T *is a local diffeomorphism around* z^* *for* $\alpha \in]0, \alpha_0[$.

Proof. By Assumptions 6.4.1, 6.4.2, and Lemma 6.3.1, $p(z)$ is continuous, therefore when z sufficiently close to z^*, $p(z)$ is in the neighbourhood of $\mathrm{Prox}_{\gamma g}(z^*)$ guaranteed by Assumption 6.4.3. Lemma 6.4.3 and chain rule tell us that

$$J_p(z) = \Big(I + \gamma\nabla^2 f\Big(p(z)\Big)\Big)^{-1} A(z),$$
$$J_{\mathrm{Prox}_{\gamma g}}(z) = \Big(I + \gamma\nabla^2 g\Big(\mathrm{Prox}_{\gamma g}(z)\Big)\Big)^{-1},$$

where $A(z)$ is defined in (6.12), so $J_T(z)$ exists and $J_T(z) = I + \alpha(J_p(z) - J_{\mathrm{Prox}_{\gamma g}}(z))$ for z sufficiently close to z^*.

For the local invertibility of T around z^*, let us show that $\det(J_T(z)) > 0$ for z sufficiently close to z^*.

First, let us define $\psi(\alpha) \coloneqq \det(I + \alpha B)$, where $B \in \mathbb{R}^{n\times n}$. Then, we know that $\psi(\alpha)$ is a polynomial of α and $\psi(0) = 1$.

Furthermore, the coefficients of this polynomial are sums of products of the entries of B. Since each entry is bounded by $\|B\|_F$ and $\|B\|_F \leq \sqrt{n}\|B\|$, we know that all the coefficients of $\psi(\alpha)$ can be bounded by some polynomial of $\|B\|$.

Now let us set

$$\begin{aligned} B &= J_p(z) - J_{\mathrm{Prox}_{\gamma g}}(z) \\ &= \Big(I + \gamma\nabla^2 f\big(p(z)\big)\Big)^{-1} A(z) - \Big(I + \gamma\nabla^2 g\big(\mathrm{Prox}_{\gamma g}(z)\big)\Big)^{-1}, \end{aligned}$$

where $A(z)$ is given in (6.12), we know that $\|B\|$ is bounded for all $z \in \mathbb{R}^{\mathrm{n}}$.

As a result, there exists $\alpha_0 > 0$ such that $\det(J_T(z)) > 0$ for all $\alpha \in (0, \alpha_0)$. □

□

Now we are ready to show the main result of this section: when α is small enough, the probability for DRS and FBS to converge to any strict saddle point of φ^γ is 0, which is also true for any strict saddle point of φ.

Lemma 6.5.4. *Let Assumptions 6.4.1 and 6.4.2 hold, then $z^* \in \mathrm{Fix}T$ if and only if $\nabla\varphi^\gamma(z^*) = \mathbf{0}$.*

Proof. This follows directly from Theorem 6.4.4. □ □

Theorem 6.5.5. *Define $Z^* = \{z^* \in \mathbb{R}^{\mathrm{n}} \mid \nabla\varphi^\gamma(z^*) = 0, \lambda_{\min}\big(\nabla^2\varphi^\gamma(z^*)\big) < 0\}$ as the set of strict saddle points of φ^γ. Under Assumptions 6.4.1, 6.4.2, and 6.4.3, if either $g = 0$ or $h = 0$, then, for sufficiently small α, we have that the set $W \coloneqq \{z^0 \in \mathbb{R}^{\mathrm{n}} : \lim_{k\to\infty} T^k z^0 \in Z^*\}$ satisfies $\mu(W) = 0$.*

Proof. Take any $z^* \in Z^*$, Lemma 6.4.3 states that $A(z^*)$ is invertible and symmetric. Also, $\nabla^2\varphi(z^*)$ is symmetric.

According to (6.18), we have

$$J_p(z^*) - J_{\mathrm{Prox}_{\gamma g}}(z^*) = -\gamma A^{-1}(z^*)\nabla^2\varphi^\gamma(z^*),$$

since $A^{-1}(z^*)\nabla^2\varphi^\gamma(z^*)$ is similar to $A^{-\frac{T}{2}}(z^*)\nabla^2\varphi^\gamma(z^*)A^{-\frac{1}{2}}(z^*)$, we know that $J_p(z^*) - J_{\mathrm{Prox}_{\gamma g}}(z^*)$ has real eigenvalues and

$$\lambda_{\max}\Big(J_p(z^*) - J_{\mathrm{Prox}_{\gamma g}}(z^*)\Big) > 0.$$

Since

$$\lambda_{\max}\Big(J_T(z^*)\Big) = 1 + \alpha\lambda_{\max}\Big(J_p(z^*) - J_{\mathrm{Prox}_{\gamma g}}(z^*)\Big),$$

we know that $Z^* \subseteq A_T^*$. Furthermore, from Lemma 6.5.3 we know that T is a local diffeomorphism around every $z^* \in Z^* \subseteq A_T^*$, therefore Theorem 6.5.2 gives $\mu(W) = 0$. □ □

Theorem 6.5.6. *Define* $X^* \coloneqq \{x^* \in \mathbb{R}^n \mid \nabla\varphi(x^*) = 0, \lambda_{\min}(\nabla^2\varphi(x^*)) < 0\}$ *as the set of strict saddle points of* φ. *Under Assumptions 6.4.1, 6.4.2, and 6.4.3, if either* $g = 0$ *or* $h = 0$, *then, the set* $V \coloneqq \{z^0 \in \mathbb{R}^n : \lim_{k\to\infty} \mathrm{Prox}_{\gamma g}(T^k z^0) \in X^*\}$ *satisfies* $\mu(V) = 0$.

Proof. Combine Theorem 6.4.8 with Theorem 6.5.5. □ □

When the objective satisfies the *strict saddle property*, i.e., the stationary points of the objective are either local minimizers or strict saddle points, we can conclude that FBS and DRS almost always converge to local minimizers of the objective whenever they converge.

Many problems in practice satisfy the strict saddle property. Examples include dictionary learning [213], simple neural networks [46], phase retrieval [212], tensor decomposition [97], and low rank matrix factorization [33].

6.6 Conclusions

In this chapter, we have constructed an envelope for DYS and established various properties of this envelope. Specifically, there are one-to-one correspondences between the global, local minimizers, critical (stationary) points and strict saddle points of this envelope and those of the original objective. Then, by the Stable-Center Manifold theorem, we have shown that the probability for FBS or DRS to converge from random starting points to strict saddle points of the envelope is zero. If the original objective also satisfies the strict saddle property, we have concluded that, whenever FBS and DRS converge, their iterates will almost always converge to local minimizers.

A limitation of this work lies in its smoothness assumptions. The construction of the envelope requires the Lipschitz differentiability of $g(x)$. Furthermore, twice differentiability of $f(x)$ at specific points is needed for the strict saddle avoidance property of FBS and DRS. It is undoubtedly interesting to investigate the possibility of weakening these assumptions in the future.

Bibliography

[1] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205. ACM, 2017.

[2] Zeyuan Allen-Zhu. Katyusha X: Practical Momentum Method for Stochastic Sum-of-Nonconvex Optimization. In *ICML*, 2018.

[3] Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pages 1110–1119, 2016.

[4] Zeyuan Allen-Zhu and Yang Yuan. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *International conference on machine learning*, pages 1080–1089, 2016.

[5] Mosek ApS. Mosek optimization toolbox for matlab. *Users Guide and Reference Manual, version*, 4, 2019.

[6] Francisco J Aragón Artacho, Jonathan M Borwein, and Matthew K Tam. Douglas–Rachford feasibility methods for matrix completion problems. *The ANZIAM J.*, 55(4):299–326, 2014.

[7] Francisco J Aragón Artacho, Jonathan M Borwein, and Matthew K Tam. Global behavior of the Douglas–Rachford method for a nonconvex feasibility problem. *J. of Global Optim.*, 65(2):309–327, 2016.

[8] H. Attouch and M. Théra. A general duality principle for the sum of two operators. *Journal of Convex Analysis*, 3(1):1–24, 1996.

[9] Hedy Attouch, Jean-Bernard Baillon, and Michel A Théra. *Variational sum of monotone operators*. Department des Sciences Mathématiques, 1994.

[10] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[11] Jean-Bernard Baillon, Ronald E. Bruck, and Simeon Reich. On the asymptotic behavior of nonexpansive mappings and semigroups in Banach spaces. *Houston Journal of Mathematics*, 4(1):1–9, 1978.

[12] G. Banjac, P. Goulart, B. Stellato, and S. Boyd. Infeasibility detection in the alternating direction method of multipliers for convex optimization. *Optimization-online.org*, 2017.

[13] Heinz H. Bauschke, Radu I. Boţ, Warren L. Hare, and Walaa M. Moursi. Attouch-Théra duality revisited: paramonotonicity and operator splitting. *Journal of Approximation Theory*, 164(8):1065–1084, 2012.

[14] Heinz H Bauschke, Jonathan M Borwein, and Adrian S Lewis. The method of cyclic projections for closed convex sets in Hilbert space. *Contemp. Math.*, 204:1–38, 1997.

[15] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York, New York, NY, 2011.

[16] Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2011.

[17] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer New York, 2nd edition, 2017.

[18] Heinz H Bauschke, Patrick L Combettes, et al. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 2011. Springer, 2017.

[19] Heinz H. Bauschke, Patrick L. Combettes, and D.Russell Luke. Finding best approximation pairs relative to two closed convex sets in Hilbert spaces. *Journal of Approximation Theory*, 127(2):178–192, 2004.

[20] Heinz H. Bauschke, Graeme R. Douglas, and Walaa M. Moursi. On a result of Pazy concerning the asymptotic behaviour of nonexpansive mappings. *J. Fixed Point Theory Appl.*, 18(2):297–307, 2016.

[21] Heinz H. Bauschke, Warren L. Hare, and Walaa M. Moursi. Generalized solutions for the sum of two maximally monotone operators. *SIAM Journal on Control and Optimization*, 52(2):1034–1047, 2014.

[22] Heinz H Bauschke, Warren L Hare, and Walaa M Moursi. Generalized solutions for the sum of two maximally monotone operators. *SIAM J. Control Optim.*, 52(2):1034–1047, 2014.

[23] Heinz H. Bauschke, Warren L. Hare, and Walaa M. Moursi. On the range of the DouglasRachford operator. *Mathematics of Operations Research*, 41(3):884–897, 2016.

[24] Heinz H Bauschke and Walaa M Moursi. The Douglas–Rachford algorithm for two (not necessarily intersecting) affine subspaces. *SIAM J. Optim.*, 26(2):968–985, 2016.

[25] Heinz H. Bauschke and Walaa M. Moursi. The Douglas-Rachford algorithm for two (not necessarily intersecting) affine subspaces. *SIAM Journal on Optimization*, 26(2):968–985, 2016.

[26] Heinz H. Bauschke and Walaa M. Moursi. On the Douglas-Rachford algorithm. *Mathematical Programming*, 164(1):263–284, Jul 2017.

[27] Heinz H. Bauschke and Walaa M. Moursi. On the Douglas–Rachford algorithm. *Mathematical Programming*, 164(1-2):263–284, 2017.

[28] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[29] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[30] Stephen R Becker, Emmanuel J Candès, and Michael C Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical programming computation*, 3(3):165, 2011.

[31] D.P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

[32] D.P. Bertsekas and J.N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Englewood Cliffs, 1989.

[33] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[34] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Nonconvex lagrangian-based optimization: monitoring schemes and global convergence. *Mathematics of Operations Research*, 43(4):1210–1232, 2018.

[35] J. M. Borwein and H. Wolkowicz. Characterizations of optimality without constraint qualification for the abstract convex program. In Monique Guignard, editor, *Optimality and Stability in Mathematical Programming*, pages 77–100. Springer Berlin Heidelberg, Berlin, Heidelberg, 1982.

[36] Jon Borwein and Henry Wolkowicz. Regularizing the abstract convex program. *Journal of Mathematical Analysis and Applications*, 83(2):495–530, 1981.

[37] Jon Borwein and Henry Wolkowicz. Regularizing the abstract convex program. *J. Math. Anal. Appl.*, 83(2):495–530, 1981.

[38] Jon M. Borwein and Henry Wolkowicz. Facial reduction for a cone-convex programming problem. *Journal of the Australian Mathematical Society*, 30(3):369–380, 1981.

[39] Jon M Borwein and Henry Wolkowicz. Facial reduction for a cone-convex programming problem. *J. Austral. Math. Soc.*, 30(3):369–380, 1981.

[40] Jonathan M Borwein and Matthew K Tam. The cyclic Douglas–Rachford method for inconsistent feasibility problems. *J. Nonlinear Convex Anal.*, 16(4):537–584, 2015.

[41] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.

[42] Kristian Bredies and Hongpeng Sun. Preconditioned Douglas-Rachford splitting methods for convex-concave saddle-point problems. *SIAM Journal on Numerical Analysis*, 53(1):421–444, 2015.

[43] Kristian Bredies and Hongpeng Sun. A proximal point analysis of the preconditioned alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 173(3):878–907, 2017.

[44] Luis M Briceño-Arias and Patrick L Combettes. A monotone+ skew splitting model for composite monotone inclusions in duality. *SIAM Journal on Optimization*, 21(4):1230–1250, 2011.

[45] Luis M Briceño-Arias and Damek Davis. Forward-backward-half forward algorithm for solving monotone inclusions. *SIAM Journal on Optimization*, 28(4):2839–2871, 2018.

[46] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with Gaussian inputs. In *International Conference on Machine Learning*, pages 605–614, 2017.

[47] Richard H Byrd, Samantha L Hansen, Jorge Nocedal, and Yoram Singer. A stochastic quasi-Newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

[48] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[49] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[50] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

[51] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[52] Liang Chen, Xudong Li, Defeng Sun, and Kim-Chuan Toh. On the equivalence of inexact proximal alm and admm for a class of convex composite programming. *Math. Program.*, 2019.

[53] Liang Chen, Defeng Sun, and Kim-Chuan Toh. A note on the convergence of ADMM for linearly constrained convex optimization problems. *Computational Optimization and Applications*, 66(2):327–343, 2017.

[54] Pengwen Chen and Albert Fannjiang. Fourier phase retrieval with a single mask by Douglas-Rachford algorithms. *Applied and Computational Harmonic Analysis*, 2016.

[55] Yuen-Lam Cheung, Simon Schurr, and Henry Wolkowicz. Preprocessing and regularization for degenerate semidefinite programs. In David H. Bailey, Heinz H. Bauschke, Peter Borwein, Frank Garvan, Michel Théra, Jon D. Vanderwerff, and Henry Wolkowicz, editors, *Computational and Analytical Mathematics*, pages 251–303. Springer New York, New York, NY, 2013.

[56] Yat Tin Chow, Tianyu Wu, and Wotao Yin. Cyclic coordinate-update algorithms for fixed-point problems: Analysis and applications. *SIAM Journal on Scientific Computing*, 39(4):A1280–A1300, 2017.

[57] Patrick L. Combettes. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53(5–6):475–504, 2004.

[58] Patrick L. Combettes. Monotone operator theory in convex optimization. *Mathematical Programming*, 170(1):177–206, 2018.

[59] Patrick L. Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In Heinz H. Bauschke, Regina S. Burachik, Patrick L. Combettes, Veit Elser, D. Russell Luke, and Henry Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer New York, New York, NY, 2011.

[60] Patrick L Combettes and Noli N Reyes. Moreaus decomposition in banach spaces. *Mathematical Programming*, 139(1-2):103–114, 2013.

[61] Patrick L Combettes and Băng C Vũ. Variable metric forward-backward splitting with applications to monotone inclusions in duality. *Optimization*, 63(9):1289–1318, 2014.

[62] Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

[63] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[64] Inc. CVX Research. CVX: Matlab software for disciplined convex programming, version 2.0. http://cvxr.com/cvx, August 2012.

[65] D. Davis. Convergence rate analysis of primal-dual splitting schemes. *SIAM Journal on Optimization*, 25(3):1912–1943, 2015.

[66] Damek Davis and Wotao Yin. Convergence rate analysis of several splitting schemes. In Roland Glowinski, Stanley Osher, and Wotao Yin, editors, *Splitting Methods in Communication, Imaging, Science and Engineering*, Chapter 4. Springer, 2016.

[67] Damek Davis and Wotao Yin. Convergence rate analysis of several splitting schemes. In Roland Glowinski, Stanley J. Osher, and Wotao Yin, editors, *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 115–163. Springer International Publishing, 2016.

[68] Damek Davis and Wotao Yin. Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions. *Math. Oper. Res.*, 42(3):783–805, 2017.

[69] Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-valued and variational Analysis*, 25(4):829–858, 2017.

[70] Etienne de Klerk, Tamás Terlaky, and Kees Roos. Self-dual embeddings. In Henry Wolkowicz, Romesh Saigal, and Lieven Vandenberghe, editors, *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, pages 111–138. Springer US, 2000.

[71] Jesus A. De Loera, Peter N. Malkin, and Pablo A. Parrilo. Computation with polynomial equations and inequalities arising in combinatorial optimization. In Jon Lee and Sven Leyffer, editors, *Mixed Integer Nonlinear Programming*, pages 447–481. Springer New York, New York, NY, 2012.

[72] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27*, pages 1646–1654, 2014.

[73] Aaron Defazio, Justin Domke, and Caetano. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning*, pages 1125–1133, January 2014.

[74] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916, 2016.

[75] DA d'Esopo. A convex programming procedure. *Naval Research Logistics Quarterly*, 6(1):33–42, 1959.

[76] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[77] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Trans. Amer. Math. Soc.*, 82:421–439, 1956.

[78] Dmitriy Drusvyatskiy and Henry Wolkowicz. The many faces of degeneracy in conic optimization. *Foundations and Trends in Optimization*, 3(2):77–170, 2017.

[79] Joseph C Dunn and S Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.

[80] J. Eckstein. *Splitting methods for monotone operators with applications to parallel optimization*. PhD thesis, MIT, 1989.

[81] J. Eckstein and D. P. Bertsekas. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1–3), 1992.

[82] Jonathan Eckstein. *Splitting Methods for Monotone Operators with Applications to Parallel Optimization*. PhD thesis, MIT, 1989.

[83] Jonathan Eckstein and Dimitri P Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

[84] Jonathan Eckstein and Masao Fukushima. Some reformulations and applications of the alternating direction method of multipliers. In W. W. Hager, D. W. Hearn, and P. M. Pardalos, editors, *Large Scale Optimization*, pages 115–134. Springer, 1994.

[85] Jonathan Eckstein and Wang Yao. Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. *Pacific Journal on Optimization*, 11(4):619–644, 2015.

[86] Jonathan Eckstein and Wang Yao. Approximate ADMM algorithms derived from lagrangian splitting. *Computational Optimization and Applications*, 68(2):363–405, 2017.

[87] Jonathan Eckstein and Wang Yao. Relative-error approximate versions of Douglas-Rachford splitting and special cases of the ADMM. *Mathematical Programming*, pages 1–28, 2017.

[88] Ernie Esser, Xiaoqun Zhang, and Tony F Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

[89] M. Fazel, T. Pong, D. Sun, and P. Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

[90] Diego Feijer and Fernando Paganini. Stability of primal–dual gradient dynamics and applications to network optimization. *Automatica*, 46(12):1974–1981, 2010.

[91] W. Fenchel. Convex cones, sets, and functions, 1953. mimeographed lecture notes.

[92] Michel Fortin and Roland Glowinski. Chapter iii on decomposition-coordination methods using an augmented lagrangian. In *Studies in Mathematics and Its Applications*, volume 15, pages 97–146. Elsevier, 1983.

[93] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[94] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.*, 2(1):17–40, 1976.

[95] Daniel Gabay. Chapter ix applications of the method of multipliers to variational inequalities. In *Studies in mathematics and its applications*, volume 15, pages 299–331. Elsevier, 1983.

[96] Carl Friedrich Gauss. Werke (in german), 9. *Göttingen: Köninglichen Gesellschaft der Wissenschaften*, 763:764, 1903.

[97] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[98] Philip E Gill, Walter Murray, Michael A Saunders, John A Tomlin, and Margaret H Wright. On projected newton barrier methods for linear programming and an equivalence to karmarkars projective method. *Mathematical programming*, 36(2):183–209, 1986.

[99] Pontus Giselsson and Stephen Boyd. Diagonal scaling in Douglas-Rachford splitting and ADMM. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pages 5033–5039. IEEE, 2014.

[100] Pontus Giselsson and Stephen Boyd. Linear convergence and metric selection for Douglas-Rachford splitting and ADMM. *IEEE Transactions on Automatic Control*, 62(2):532–544, 2017.

[101] Pontus Giselsson and Mattias Fält. Envelope functions: Unifications and further properties. *Journal of Optimization Theory and Applications*, 178(3):673–698, 2018.

[102] R. Glowinski. *Numerical methods for nonlinear variational problems*. Springer, 1984.

[103] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *Revue Française d'Automatique, Informatique, Recherche Opérationnelle. Analyse Numérique*, 9(2):41–76, 1975.

[104] Roland Glowinski and A Marrocco. On the solution of a class of non linear dirichlet problems by a penalty-duality method and finite elements of order one. In *Optimization Techniques IFIP Technical Conference*, pages 327–333. Springer, 1975.

[105] Alon Gonen, Francesco Orabona, and Shai Shalev-Shwartz. Solving ridge regression using sketched preconditioned SVRG. In *International Conference on Machine Learning*, pages 1397–1405, 2016.

[106] Robert Gower, Donald Goldfarb, and Peter Richtárik. Stochastic block BFGS: squeezing more curvature out of data. In *International Conference on Machine Learning*, pages 1869–1878, 2016.

[107] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. http://stanford.edu/~boyd/graph_dcp.html.

[108] Ke Guo, Deren Han, and Xiaoming Yuan. Convergence analysis of Douglas-Rachford splitting method for strongly+ weakly convex programming. *SIAM Journal on Numerical Analysis*, 55(4):1549–1577, 2017.

[109] Robert Hannah, Fei Feng, and Wotao Yin. A2BCD: An asynchronous accelerated block coordinate descent algorithm with optimal complexity. *arXiv preprint arXiv:1803.05578*, 2018.

[110] Robert Hannah, Yanli Liu, Daniel O'Connor, and Wotao Yin. Breaking the span assumption yields fast finite-sum minimization. In *Advances in Neural Information Processing Systems*, pages 2314–2323, 2018.

[111] Per Christian Hansen and Jakob Sauer Jørgensen. Air tools ii: algebraic iterative reconstruction methods, improved implementation. *Numerical Algorithms*, 79(1):107–137, 2018.

[112] Bingsheng He and Xiaoming Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1):119–149, 2012.

[113] Clifford Hildreth. A quadratic programming procedure. *Naval research logistics quarterly*, 4(1):79–85, 1957.

[114] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323, 2013.

[115] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[116] R. B. Kellogg. A nonlinear alternating direction method. *Math. Comput.*, 23(105):23–27, 1969.

[117] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

[118] K So Kretschmer. Programmes in paired spaces. *Canad. J. Math.*, 13:221–238, 1961.

[119] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

[120] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

[121] Lihua Lei and Michael Jordan. Less than a single pass: stochastically controlled stochastic gradient. In *Artificial Intelligence and Statistics*, pages 148–156, 2017.

[122] Elizaveta Levina and Peter Bickel. The earth mover's distance is the mallows distance: Some insights from statistics. In *null*, page 251. IEEE, 2001.

[123] Evgeny S Levitin and Boris T Polyak. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 6(5):1–50, 1966.

[124] Guoyin Li, Tianxiang Liu, and Ting Kei Pong. Peaceman-Rachford splitting for a class of nonconvex optimization problems. *Computational Optimization and Applications*, 68(2):407–436, 2017.

[125] Guoyin Li and Ting Kei Pong. Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Mathematical programming*, 159(1-2):371–401, 2016.

[126] Min Li, Li-Zhi Liao, and Xiaoming Yuan. Inexact alternating direction methods of multipliers with logarithmic-quadratic proximal regularization. *Journal of Optimization Theory and Applications*, 159(2):412–436, 2013.

[127] WUCHEN Li, ERNEST K Ryu, STANLEY Osher, WOTAO Yin, and WILFRID Gangbo. A parallel method for earth mover's distance. *UCLA Comput. Appl. Math. Pub.(CAM) Rep*, pages 17–12, 2017.

[128] Sophus Lie. Theorie der transformationsgruppen abschn. 3. *Theorie der Transformationsgruppen*, 1893.

[129] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.

[130] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. An inexact variable metric proximal point algorithm for generic quasi-newton acceleration. *arXiv preprint arXiv:1610.00960*, 2016.

[131] Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated proximal coordinate gradient method. In *Advances in Neural Information Processing Systems*, pages 3059–3067, 2014.

[132] Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4):2244–2273, 2015.

[133] Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[134] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):208–220, 2013.

[135] Minghui Liu and Gábor Pataki. Exact duals and short certificates of infeasibility and weak infeasibility in conic linear programming. *Mathematical Programming*, pages 1–46, 2017.

[136] Yanli Liu, Fei Feng, and Wotao Yin. Acceleration of svrg and katyusha x by inexact preconditioning. In *International Conference on Machine Learning*, pages 4003–4012, 2019.

[137] Yanli Liu, Ernest K Ryu, and Wotao Yin. A new use of Douglas–Rachford splittingfor identifying infeasible, unbounded, and pathological conic programs. *Math. Program.*, 177(1–2):225–253, 2019.

[138] Yanli Liu, Yunbei Xu, and Wotao Yin. Acceleration of primal-dual methods by preconditioning and simple subproblem procedures. *arXiv preprint arXiv:1811.08937*, 2018.

[139] Yanli Liu and Wotao Yin. An envelope for davis–yin splitting and strict saddle-point avoidance. *Journal of Optimization Theory and Applications*, 181(2):567–587, 2019.

[140] Jesus A. De Loera, Peter N. Malkin, and Pablo A. Parrilo. Computation with polynomial equations and inequalities arising in combinatorial optimization. In *Mixed Integer Nonlinear Programming*, pages 447–481. Springer, New York, NY, 2012.

[141] J. Lofberg. Pre- and post-processing sum-of-squares programs in practice. *IEEE Transactions on Automatic Control*, 54(5):1007–1011, 2009.

[142] J. Lofberg. Pre- and post-processing sum-of-squares programs in practice. *IEEE Trans. Autom. Control*, 54(5):1007–1011, 2009.

[143] Bruno F. Lourenço, Masakazu Muramatsu, and Takashi Tsuchiya. Solving SDP completely with an interior point oracle. *arXiv:1507.08065 [math]*, 2015.

[144] Bruno F Lourenço, Masakazu Muramatsu, and Takashi Tsuchiya. Solving SDP completely with an interior point oracle. *arXiv preprint arXiv:1507.08065*, 2015.

[145] David G. Luenberger and Yinyu Ye. Conic linear programming. In *Linear and Nonlinear Programming*, number 228 in International Series in Operations Research & Management Science, pages 149–176. Springer International Publishing, 2016.

[146] Z.-Q. Luo, J. F. Sturm, and S. Zhang. Conic convex programming and self-dual embedding. *Optimization Methods and Software*, 14(3):169–218, 2000.

[147] Z-Q Luo, Jos Fredrik Sturm, and Shuzhong Zhang. Conic convex programming and self-dual embedding. *Optim. Methods Softw.*, 14(3):169–218, 2000.

[148] Zhi-Quan Luo, Jos F Sturm, and Shuzhong Zhang. Duality results for conic convex programming. *Econometric Institute, Erasmus University Rotterdam, The Netherlands, Technical Report 9719/A*, 1997.

[149] Zhi-Quan Luo, Jos F Sturm, and Shuzhong Zhang. Duality results for conic convex programming. Technical report, Erasmus University Rotterdam, Econometric Institute, 1997.

[150] Julien Mairal. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pages 783–791, February 2013.

[151] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *2009 IEEE 12th international conference on computer vision*, pages 2272–2279. IEEE, 2009.

[152] B Mercier. *Inéquations Variationnelles de la Mécanique (Publications Mathématiques d'Orsay, no. 80.01)*. Orsay, France: Université de Paris-XI, 1980.

[153] Ludovic Métivier, Romain Brossier, Quentin Mérigot, Edouard Oudet, and Jean Virieux. Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion. *Geophysical Supplements to the Monthly Notices of the Royal Astronomical Society*, 205(1):345–377, 2016.

[154] Aryan Mokhtari, Mark Eisen, and Alejandro Ribeiro. IQN: An incremental quasi-Newton method with local superlinear convergence rate. *SIAM Journal on Optimization*, 28(2):1670–1698, 2018.

[155] Philipp Moritz, Robert Nishihara, and Michael Jordan. A linearly-convergent stochastic L-BFGS algorithm. In *Artificial Intelligence and Statistics*, pages 249–258, 2016.

[156] ApS Mosek. The MOSEK optimization toolbox for matlab manual. *Version 7.1 (Revision 28)*, page 17, 2015.

[157] Walaa M. Moursi. *The Douglas–Rachford Operator in the Possibly Inconsistent Case: Static Properties and Dynamic Behaviour*. PhD thesis, University of British Columbia, 2017.

[158] Walaa M Moursi. The forward-backward algorithm and the normal problem. *Journal of Optimization Theory and Application*, 176(3):605–624, 2018.

[159] Arkadiĭ Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[160] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[161] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. Siam, 1994.

[162] Yurii Nesterov, Michael J Todd, and Yinyu Ye. Infeasible-start primal-dual methods and infeasibility detectors for nonlinear programming problems. *Mathematical Programming*, 84(2):227–267, 1999.

[163] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547, 1983.

[164] Michael K Ng, Fan Wang, and Xiaoming Yuan. Inexact alternating direction methods for image recovery. *SIAM Journal on Scientific Computing*, 33(4):1643–1668, 2011.

[165] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621, 2017.

[166] Gurobi Optimization. Gurobi optimizer reference manual; gurobi optimization. *Inc.: Houston, TX, USA*, 2016.

[167] Brendan Odonoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

[168] Gregory B Passty. Ergodic convergence to a zero of the sum of monotone operators in hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, 1979.

[169] Gábor Pataki. A simple derivation of a facial reduction algorithm and extended dual systems. *Columbia University, Technical report*, 2000.

[170] Gábor Pataki. A simple derivation of a facial reduction algorithm and extended dual systems. Technical report, Columbia University, 2000.

[171] Panagiotis Patrinos, Lorenzo Stella, and Alberto Bemporad. Douglas-Rachford splitting: Complexity estimates and accelerated variants. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pages 4234–4239. IEEE, 2014.

[172] A. Pazy. Asymptotic behavior of contractions in Hilbert space. *Israel Journal of Mathematics*, 9(2):235–240, 1971.

[173] D. W. Peaceman and H. H. Rachford. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Ind. Appl. Math.*, 3(1), 1955.

[174] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *ICCV*, volume 9, pages 460–467, 2009.

[175] Frank Permenter, Henrik A Friberg, and Erling D Andersen. Solving conic optimization problems via self-dual embedding and facial reduction: a unified approach. *SIAM Journal on Optimization*, 27(3):1257–1282, 2017.

[176] Frank Permenter and Pablo Parrilo. Partial facial reduction: Simplified, equivalent SDPs via approximations of the PSD cone. *arXiv:1408.4685*, 2014.

[177] Frank Permenter and Pablo Parrilo. Partial facial reduction: simplified, equivalent SDPs via approximations of the PSD cone. *Mathematical Programming*, 171(1–2):1–54, 2018.

[178] Frank Permenter and Pablo A. Parrilo. Basis selection for SOS programs via facial reduction and polyhedral approximations. In *53rd IEEE Conference on Decision and Control, CDC 2014*, pages 6615–6620, 2014.

[179] Thomas Pock and Antonin Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1762–1769. IEEE, 2011.

[180] Thomas Pock, Daniel Cremers, Horst Bischof, and Antonin Chambolle. An algorithm for minimizing the Mumford-Shah functional. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1133–1140. IEEE, 2009.

[181] Boris T Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1, 1987.

[182] Michael JD Powell. On search directions for minimization algorithms. *Mathematical programming*, 4(1):193–201, 1973.

[183] Arvind U. Raghunathan and Stefano Di Cairano. Infeasibility detection in alternating direction method of multipliers for convex quadratic programs. In *2014 IEEE 53rd Annual Conference on Decision and Control (CDC)*, pages 5819–5824. 2014.

[184] Motakuri V. Ramana, Levent Tunçel, and Henry Wolkowicz. Strong duality for semidefinite programming. *SIAM J. Optim.*, 7(3):641–662, 1997.

[185] Julian Rasch and Antonin Chambolle. Inexact first-order primal-dual algorithms. *arXiv preprint arXiv:1803.10576*, 2018.

[186] Julian P. Revalski and Michel Théra. Generalized sums of monotone operators. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 329(11):979–984, 1999.

[187] Julian P. Revalski and Michel Théra. Enlargements and sums of monotone operators. *Nonlinear Analysis: Theory, Methods & Applications*, 48(4):505–519, 2002.

[188] Lewis Fry Richardson. Ix. the approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Phil. Trans. R. Soc. Lond. A*, 210(459-470):307–357, 1911.

[189] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[190] R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970.

[191] R. Tyrrell Rockafellar. *Conjugate Duality and Optimization*. Society for Industrial and Applied Mathematics, 1974.

[192] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[193] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[194] R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[195] R.T. Rockafellar. *Conjugate Duality and Optimization*. Society for Industrial and Applied Mathematics, 1974.

[196] Anton Rodomanov and Dmitry Kropotov. A superlinearly-convergent proximal Newton-type method for the optimization of finite sums. In *International Conference on Machine Learning*, pages 2597–2605, 2016.

[197] Nicolas L. Roux, Mark Schmidt, and Francis R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems 25*, pages 2663–2671. Curran Associates, Inc., 2012.

[198] Ernest K Ryu. Cosmic divergence, weak cosmic convergence, and fixed points at infinity. *Journal of Fixed Point Theory and Applications*, 2018.

[199] Ernest K Ryu, Yanli Liu, and Wotao Yin. Douglas–rachford splitting and admm for pathological convex optimization. *Computational Optimization and Applications*, 74(3):747–778, 2019.

[200] Youcef Saad. *Numerical Methods for Large Eigenvalue Problems*. Manchester University Press, 1992.

[201] Yousef Saad. *Iterative Methods for Sparse Linear Systems*, volume 82. SIAM, 2003.

[202] Shai Shalev-Shwartz. SDCA without duality, regularization, and individual convexity. In *International Conference on Machine Learning*, pages 747–754, 2016.

[203] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *J. Mach. Learn. Res.*, 14(1):567–599, February 2013.

[204] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1-2):105–145, January 2016.

[205] Naum Zuselevich Shor. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012.

[206] Michael Shub. *Global Stability of Dynamical Systems*. Springer Science & Business Media, 2013.

[207] Emil Y Sidky, Jakob H Jørgensen, and Xiaochuan Pan. Convex optimization problem prototyping for image reconstruction in computed tomography with the chambolle–pock algorithm. *Physics in Medicine & Biology*, 57(10):3065, 2012.

[208] Lorenzo Stella. *Proximal Envelopes: Smooth Optimization Algorithms for Nonsmooth Problems*. PhD thesis, IMT School for Advanced Studies Lucca, Lucca, Italy, 2017.

[209] Lorenzo Stella, Andreas Themelis, and Panagiotis Patrinos. Forward-backward quasi-Newton methods for nonsmooth optimization problems. *Computational Optimization and Applications*, 67(3):443–487, 2017.

[210] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. OSQP: An operator splitting solver for quadratic programs. *arXiv preprint arXiv:1711.08013*, 2017.

[211] Jos F Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization methods and software*, 11(1-4):625–653, 1999.

[212] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 2379–2383. IEEE, 2016.

[213] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

[214] B. F. Svaiter. On weak convergence of the Douglas–Rachford method. *SIAM Journal on Control and Optimization*, 49(1):280–287, 2011.

[215] Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone line-search algorithms. *arXiv preprint arXiv:1606.06256*, 2016.

[216] Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Douglas-Rachford splitting and ADMM for nonconvex optimization: new convergence results and accelerated versions. *arXiv preprint arXiv:1709.05747*, 2017.

[217] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[218] Paul Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

[219] Levent Tunçel and Henry Wolkowicz. Strong duality and minimal representations for cone optimization. *Comput. Optim. Appl.*, 53(2):619–648, 2012.

[220] Reha H Tütüncü, Kim-Chuan Toh, and Michael J Todd. Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical programming*, 95(2):189–217, 2003.

[221] Tuomo Valkonen. A primal–dual hybrid gradient method for nonlinear operators with applications to mri. *Inverse Problems*, 30(5):055012, 2014.

[222] Robert J Vanderbei and David F Shanno. An interior-point algorithm for nonconvex nonlinear programming. *Computational Optimization and Applications*, 13(1-3):231–252, 1999.

[223] Bng Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

[224] Bng Công Vũ. A variable metric extension of the forward–backward–forward algorithm for monotone operators. *Numerical Functional Analysis and Optimization*, 34(9):1050–1065, 2013.

[225] Hayato Waki. How to generate weakly infeasible semidefinite programs via Lasserre's relaxations for polynomial optimization. *Optimization Letters*, 6(8):1883–1896, 2012.

[226] Hayato Waki and Masakazu Muramatsu. A facial reduction algorithm for finding sparse sos representations. *Operations Research Letters*, 38(5):361 – 365, 2010.

[227] Hayato Waki and Masakazu Muramatsu. Facial reduction algorithms for conic optimization problems. *Journal of Optimization Theory and Applications*, 158(1):188–215, 2013.

[228] Hayato Waki and Masakazu Muramatsu. Facial reduction algorithms for conic optimization problems. *J. Optim. Theory Appl.*, 158(1):188–215, 2013.

[229] Hayato Waki, Maho Nakata, and Masakazu Muramatsu. Strange behaviors of interior-point methods for solving semidefinite programming problems in polynomial optimization. *Computational Optimization and Applications*, 53(3):823–844, 2012.

[230] Xiaoyu Wang, Xiao Wang, and Ya-xiang Yuan. Stochastic proximal quasi-Newton methods for non-convex composite optimization. *Optimization Methods and Software*, pages 1–27, 2018.

[231] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324*, 2015.

[232] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[233] Yangyang Xu. Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming. *SIAM Journal on Optimization*, 27(3):1459–1484, 2017.

[234] Maretsugu Yamasaki. Some generalizations of duality theorems in math. program. problems. *Math. J. Okayama Univ.*, 14:69–81, 1969.

[235] Ming Yan. A new primal–dual algorithm for minimizing the sum of three functions with a linear operator. *Journal of Scientific Computing*, 76(3):1698–1717, 2018.

[236] Ming Yan and Wotao Yin. Self equivalence of the alternating direction method of multipliers. In Roland Glowinski, Stanley Osher, and Wotao Yin, editors, *Splitting Methods in Communication, Imaging, Science and Engineering*, pages 165–194. Springer, 2016.

[237] Ming Yan and Wotao Yin. Self equivalence of the alternating direction method of multipliers. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 165–194. Springer, 2016.

[238] Ming Yan and Wotao Yin. Self equivalence of the alternating direction method of multipliers. In Roland Glowinski, Stanley J. Osher, and Wotao Yin, editors, *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 165–194. Springer International Publishing, 2016.

[239] Yinyu Ye. Linear conic programming. *Manuscript. Stanford University, Stanford, CA*, 2004.

[240] Yinyu Ye, Michael J. Todd, and Shinji Mizuno. An $O(\sqrt{n}L)$-iteration homogeneous and self-dual linear programming algorithm. *Mathematics of Operations Research*, 19(1):53–67, 1994.

[241] Akiko Yoshise. Complementarity problems over symmetric cones: A survey of recent developments in several aspects. In Miguel F. Anjos and Jean B. Lasserre, editors, *Handbook on Semidefinite, Conic and Polynomial Optimization*, number 166 in International Series in Operations Research & Management Science, pages 339–375. Springer US, 2012.

[242] Fuzhen Zhang and Qingling Zhang. Eigenvalue inequalities for matrix product. *IEEE Transactions on Automatic Control*, 51(9):1506–1509, 2006.

[243] Xiaoqun Zhang, Martin Burger, and Stanley Osher. A unified primal-dual algorithm framework based on bregman iteration. *Journal of Scientific Computing*, 46(1):20–46, 2011.

[244] Mingqiang Zhu and Tony Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. *UCLA CAM Report*, 34, 2008.

[245] Yuzixuan Zhu and Gábor Pataki Quoc Tran-Dinh. Sieve-SDP: a simple facial reduction algorithm to preprocess semidefinite programs. *Mathematical Programming Computation*, 11(3):503–586, 2019.