Sparse Imaging by Nonconvex and Nonsmooth Minimizations: Analyses and Algorithms

Dissertation

zur Erlangung des akademischen Grades eines Doktor der Naturwissenschaften
an der Karl-Franzens-Universität Graz

vorgelegt von

M.Phil. Tao Wu

am Institut für Mathematik und wissenschaftliches Rechnen

Erstbegutachter:
Prof. Dr. Michael Hintermüller

Zweitbegutachter:
Prof. Dr. Wotao Yin

October 2014

Abstract

Sparsity, in a general sense, plays a vital role in modern signal and image processing. This thesis is devoted to nonconvex and nonsmooth minimization approaches to sparsity-based image processing, which splits into three major parts.

In the first part of the thesis, a nonconvex minimization model for restoring a gradient-sparse image is introduced which contains the ℓ^q-“norm”, $q \in (0, 1)$, of the gradient of the underlying image as a regularization. Hence, such a regularization term represents a nonconvex compromise between the minimization of the support of the reconstruction and the classical convex total-variation model. In our work, for the ℓ^q-norm based models in the discrete setting, existence of a minimizer is proved, and a Newton-type solution algorithm is introduced and its global as well as local superlinear convergence toward a stationary point of a locally regularized version of the problem is established. The potential nonpositive definiteness of the Hessian of the objective during the iteration is handled by a trust-region based regularization scheme. The performance of the new algorithm is also studied by means of a series of numerical tests. We also generalize our approach to the particular ℓ^q-minimization model to a wide range of sparsity-promoting models with concave priors, which finds interesting applications beyond image processing in, e.g., machine learning and optimal control of partial differential equations.

In the second part, a novel bilevel optimization framework is proposed for blind deconvolution, where both the underlying point spread function, which parameterizes the convolution operator, and the source image need to be identified. The minimization of a total-variation model is formulated as the lower-level problem, as is typically done in non-blind image deconvolution. The upper-level objective takes into account additional statistical information depending on the particular imaging modality. Bilevel problems of such type are investigated systematically in our work. Analytical properties of the lower-level solution mapping are established based on Robinson’s strong regularity condition. Furthermore, several stationarity conditions are derived from the variational geometry induced by the lower-level problem. Numerically, a projected-gradient-type method is employed to obtain a Clarke-type stationary point and its convergence properties are analyzed. We also implement an efficient version of the proposed algorithm and test it through the experiments on point spread function calibration and multiframe blind deconvolution.

The last part of the thesis concerns the so-called robust principal component pursuit (RPCP), which refers to a decomposition of a data matrix into a low-rank component and a sparse component. In our work, instead of invoking a convex-relaxation model based on the nuclear norm and the ℓ^1-norm as is typically done in this context, RPCP is solved by considering a least-squares problem subject to rank and cardinality constraints. An inexact alternating minimization scheme, with guaranteed global convergence, is employed to solve the resulting constrained minimization problem. In particular, the low-rank matrix subproblem is resolved inexactly by a tailored Riemannian optimization technique, which favorably avoids singular value decompositions in full dimension. For the overall method, a corresponding q-linear convergence theory is established. Our numerical experiments show that the newly proposed method compares competitively with a popular convex-relaxation based approach.

Zusammenfassung

Diese Dissertation widmet sich nichtkonvexen und nichtglatten Minimierungsproblemen in der auf dünner Besetztheit basierten mathematischen Bildverarbeitung. Die Arbeit ist in drei Teile gegliedert.

Im ersten Teil der Arbeit wird ein nichtkonvexes Minimierungsmodell zur Rekonstruktion eines Bildes mit dünnbesetzter Gradientenstruktur eingeführt. Dieses Modell beruht auf einer Regularisierung mittels der ℓ^q-Quasinorm (mit $q \in (0,1)$) des Gradienten des zugrundeliegenden Bildes. Dieser Regularisierungsanteil repräsentiert einen nichtkonvexen Kompromiss zwischen der Minimierung des Trägers des Gradienten und des klassischen konvexen Modells der Regularisierung mittels totaler Variation (TV-Modell). In unserer Arbeit wird die Existenz einer Minimalstelle für das diskrete ℓ^q-Modell nachgewiesen. Ein Newton-ähnlicher Lösungsalgorithmus wird für eine regularisierte Variante des Problems eingeführt. Für diesen Algorithmus wird sowohl die globale als auch die lokale superlineare Konvergenz zu einem stationären Punkt nachgewiesen. Zur Stabilisierung aufgrund einer eventuell nicht positiv-definiten Hesse-Matrix des Zielfunktionals während der Iterationen wird ein *Trust-Region*-Verfahren verwendet. Der neue Algorithmus wird anhand numerischer Tests studiert und validiert. Anschließend wird das ℓ^q-Minimierungsmodell auf weitere Anwendungen mit dünnbesetzter Information ausgedehnt. Auf diese Weise führt der Ansatz–abgesehen von der Bildverarbeitung–auf weitere interessante Bereiche wie zum Beispiel maschinelles Lernen und optimale Steuerung von partiellen Differentialgleichungen.

Im zweiten Teil der Arbeit wird eine neuer zweistufiger Optimierungsansatz für blinde Entfaltung vorgestellt. Dabei muss sowohl der zugrundeliegende Konvolutionskern, welcher die Faltung parametrisiert, als auch eine Rekonstruktion des Bildes gefunden werden. Das TV-Modell tritt als untergeordnetes Optimierungsproblem auf, welches durch den Faltungskern parametrisiert ist. Das Zielfunktional des übergeordneten Problems berücksichtigt zusätzliche statistische Informationen, welche vom speziellen bildgebenden Verfahren abhängen. Die resultierenden zweistufigen Probleme werden systematisch untersucht. So erhält man aufgrund der starken Regularität nach Robinson analytische Eigenschaften der Lösungsabbildung des untergeordneten Problems. Weiter werden anhand verschiedener Eigenschaften der nichtglatten Geometrie des untergeordneten Problems verschiedene Stationaritätsbedingungen hergeleitet. Für die numerische Behandlung wird ein projiziertes gradienten-ähnliches Abstiegsverfahren zur Bestimmung eines Clarke-stationären Punkts entwickelt und analysiert. Daneben wird eine effiziente Variante des Algorithmus implementiert und anhand von Beispielen zur Kalibrierung des Faltungskerns und zur Entfaltung im Falle multipler Datenbilder getestet.

Der letzte Teil der Dissertation betrifft die so genannte robuste Bestimmung von Hauptkomponenten gegebener Daten. Dabei erfolgt eine Zerlegung der Datenmatrix in eine sogenannte Niedrig-Rang-Matrix und eine dünnbesetzte Matrix. Anstelle eines üblicherweise verwendeten konvexen Relaxationsmodells, welches auf der Spurnorm und der ℓ^1-Norm basiert, wird in unserer Arbeit das Problem unter Hinzuziehen einer Methode der kleinsten Quadrate mit Rang- und Besetztheitsrestriktion gelöst. Für die Lösung des resultierenden restringierten Minimierungsproblems wird ein inexaktes alternierendes Minimierungsschema, welches globale Konvergenz garantiert, angewendet. Im Speziellen wird das Teilproblem zur Berechnung der Niedrig-Rang-Komponente unter Verwendung einer Riemann'schen Optimierungstechnik inexakt gelöst. Dadurch wird eine Singulärwertzerlegung im hochdimensionalen Raum vermieden. Für das Gesamtverfahren wird q-lineare Konvergenz bewiesen. Unsere numerischen Berechnungen zeigen, dass die neue Methode im Vergleich zur gängigen konvexen Relaxationstechnik sehr konkurrenzfähig ist.

Acknowledgements

First of all, I would like to thank Prof. Michael Hintermüller for being an excellent Ph.D. advisor. His passion and insights on scientific research have provided me tremendous encouragement and support during the past four years. It has been my great pleasure to work with him, and hopefully our collaboration will continue in the future.

I have been fortunate to work with many talented and helpful colleagues within the START- and SFB-Projects in Graz and the math institute at KFU, from whom I have benefited so much both personally and academically. Among many others, I particularly thank Martin Kanitsar, Carlos Rautenberg, Andreas Langer, Martin Holler, and Matthias Schlögl for their friendships.

I gratefully acknowledge financial support from FWF for my research activities at large and several fundings from SIAM and NAWI Graz for my conference travels.

I also thank Prof. Wotao Yin for serving on my thesis committee, and Prof. Raymond Chan for his guidance during my early academic career in Hong Kong.

Lastly, I would like to dedicate this thesis to my parents for their love and support over my entire life.

Contents

1 Motivation and organization of the thesis **7**

2 Nonconvex TV^q-models in image restoration: analysis, algorithm, numerics, and generalizations **11**

2.1 Introduction . . . 11

2.2 Preliminaries . . . 13

2.2.1 Functional analytic aspects of total-variation models . . . 14

2.2.2 Semismooth Newton method . . . 19

2.2.3 Globalization strategies for nonconvex minimizations . . . 21

2.3 TV^q-model and its Huberization . . . 23

2.4 Primal-dual Newton method . . . 27

2.4.1 Regularized Newton via reweighted Euler-Lagrange equation . . . 27

2.4.2 Infeasible Newton technique . . . 28

2.4.3 Superlinear convergence by adaptive regularization . . . 30

2.5 Numerical experiments . . . 40

2.6 A smoothing scheme and the consistency result . . . 47

2.7 A note on TV^q-models in function space . . . 49

2.8 Beyond TV^q — variational models with concave sparsity-promoting priors . . . 51

2.8.1 Variational models with concave priors . . . 52

2.8.2 A superlinearly convergent regularized Newton scheme . . . 53

2.8.3 Selected applications . . . 62

3 Blind deconvolution via bilevel optimization **69**

3.1 Introduction . . . 69

3.2 Preliminaries on mathematical programs with equilibrium constraints (MPECs) . 71

3.2.1 Lower-level problem: strong regularity condition and an implicit map . . 71

3.2.2 Elements in Mordukhovich's generalized differential calculus . . . 72

3.2.3 Mordukhovich-type stationarity condition for MPECs . . . 74

3.3 A bilevel optimization model for blind deconvolution . . . 76

3.4 Solution mapping for lower-level problem: existence, continuity, and differentiability 78

3.5 Stationarity conditions for bilevel optimization 82
3.6 Hybrid projected gradient method 84
3.6.1 Differentiability given strict complementarity 84
3.6.2 Local smoothing at a non-differentiable point 87
3.6.3 Hybrid projected gradient method 90
3.7 Numerical experiments 94
3.7.1 Implementation issues 95
3.7.2 Calibration of point spread functions 98
3.7.3 Multiframe blind deconvolution 99

4 Robust principal component pursuit: a Riemannian optimization approach 105
4.1 Introduction 105
4.2 Preliminaries on Riemannian optimization 107
4.3 Robust principal component pursuit 113
4.4 Alternating minimization on matrix manifolds 115
4.4.1 Alternating minimization scheme and its convergence property 115
4.4.2 Sparse matrix (B-)subproblem 116
4.4.3 Low-rank matrix (A-)subproblem 117
4.4.4 Alternating minimization scheme with trimming 128
4.5 Numerical experiments 129
4.5.1 Numerical behavior 129
4.5.2 Comparison with an augmented Lagrangian method 130
4.5.3 Application to background-foreground separation of surveillance video . . 132
4.6 Appendix on local convergence of an alternating projection method. 134

5 Conclusion and outlook 137

Bibliography 138

Chapter 1

Motivation and organization of the thesis

Sparsity, possibly varying in form from case to case, plays a vital role in modern signal and image processing. Digital images are commonly sparse under certain linear transforms, i.e. a fraction of the transformed coefficients are dominating the rest in magnitude. The choice of the sparsifying linear transform may depend on the underlying image content and the goal of an image processing task. For example, a piecewise constant image is obviously sparse under a gradient transform. The (block) discrete cosine transform well sparsifies a common photograph from a camera, which leads to the success of JPEG-format image compression. The format JPEG-2000, a successor of JPEG, rather relies on discrete wavelet transforms and yields superior compression performance over JPEG. As is expected, sparsity-based image processing, or *sparse imaging* in short, reaches far beyond image compression. Once the sparsity of the underlying image is acknowledged as our a priori knowledge, it serves as a proper regularization of the solutions for many imaging-related, most likely ill-posed, inverse problems. Such inverse problems include image reconstruction, denoising, deblurring, inpainting, superresolution, and segmentation, to name a few. In certain medical applications such as computed tomography (CT) and magnetic resonance imaging (MRI), data measurements in respective transform-domains can be severely inadequate in amount (i.e. strongly undersampled) due to physical and physiological constraints. In such scenarios, utilization of a sparsity prior in a variational image reconstruction approach would compensate, to a certain extent, the loss of information and hence trigger faster image acquisition without degradation of image quality.

Beyond image processing, sparsity is also crucial for processing more general datasets of high dimensions in large scales. In particular, the low-rank property of a matrix or tensor can be viewed as the sparsity with respect to singular values, which finds profound applications in machine learning and data mining. For instance, a low-rank matrix or tensor may arise from a low-degree statistical model of a random process, a low-dimensional manifold embedding of high-dimensional data, or a low-order approximation of a linear operator on an infinite dimensional function space.

Motivated by the sparsity in a general sense, the present thesis concerns variational methods for obtaining sparse solutions from given data. In this regard, we propose novel nonconvex or/and nonsmooth minimization models for three different problems in image/video processing, respectively. Each problem constitutes an individual chapter; see chapters 2–4. We will observe from our numerical experiments that the newly proposed nonconvex or/and nonsmooth models, when properly utilized, indeed yield visible improvements on either quality of the solutions or computational time. Yet, nonconvex and nonsmooth minimizations are very challenging both analytically and numerically. Our goal in each problem is to investigate existence of solutions for the respective variational model, characterize optimality conditions, and devise an efficient numerical solver with complete convergence analysis.

More specifically, in chapter 2 we propose a nonconvex ℓ^q-type ($0 < q < 1$) functional for promoting sparsity of restored images. It is known, e.g. in compressed sensing, that the most straightforward quantification for the sparsity of a vector is the ℓ^0-norm, which counts the number of its nonzero entries. Nevertheless, minimization with the ℓ^0-norm is often an NP-hard combinatorial problem [Nat95]. As a comprise, a vast amount of the literature resorts to the convex ℓ^1-minimization, since the ℓ^1-norm is the convex relaxation of the ℓ^0-norm on the closed unit ball; see [CDS01, BDE09] for an overview. In image processing, one typically intends to keep edges in the solution image and hence minimizes with the total-variation (TV) norm, which amounts to the ℓ^1-norm of the image gradient in a discrete setting; see [ROF92] and its related works. More recently, there is evidence that the nonconvex ℓ^q-norm based models better preserve sparsity of the underlying solution than the ℓ^1-norm based model, which may, e.g., favorably reduce the amount of data required in image acquisition; see [Nik02, Cha07b, CY08, NNZC08]. Numerical solution for the ℓ^q-model represents a nonconvex and non-Lipschitz minimization, known to be far more challenging than convex and Lipschitz ℓ^1-minimization. In chapter 2 of the present thesis, this challenge is tackled systematically. For the nonconvex ℓ^q-model in the discrete setting, existence of a minimizer is proved, and a Newton-type solution algorithm is introduced and its global as well as local superlinear convergence toward a stationary point of a locally regularized version of the problem is established. The potential nonpositive definiteness of the Hessian of the objective during the iteration is handled by a trust-region based regularization scheme. The performance of the new algorithm is also studied by means of a series of numerical tests. It turns out that our approach to the particular ℓ^q-minimization model can be generalized to a wide range of sparsity-promoting models with *concave priors*, which finds interesting applications beyond image processing in, e.g., machine learning and optimal control of partial differential equations.

Once it is accepted that the sparsity-based variational method faithfully restores the original image, more comprehensive modeling of unknown parameters, in addition to the underlying image itself, becomes an interesting question. Such unknown parameters can be related to either sparsity prior(s) or image acquisition; see, e.g., [KP13, DlRS13]. One specific paradigm to approach this problem is *bilevel optimization.* In this context, in an upper level, in contrast to sparsity-based image restoration in a lower level, one can, for instance, minimize an energy functional, which selects the best restored image(s) according to a certain statistical criterion. In chapter 3, we investigate a particular problem of such type, namely *blind deconvolution*, using the bilevel optimization approach. Image blur is widely encountered in astronomy, microscopes, tomographic imaging, etc; see e.g. [KH96a, KH96b, CE07] and the references therein. In many situations, the blurring operator, often modeled by the convolution with a single point spread function provided that the blurring is shift-invariant, is not available beforehand and needs to be identified together with the underlying image. In this work, we restrict our sparsity prior in the lower-level problem to be the (convex) total variation only. We emphasize, however, that the overall bilevel problem represents a nonconvex and nonsmooth minimization. Moreover, the constraint that arises from solving the lower-level problem is typically characterized as a set-valued equation or a nonsmooth equation, which renders the classical Karush-Kuhn-Tucker (KKT) theory inapplicable for deriving optimality conditions of our bilevel optimization. Instead, we apply Mordukhovich's generalized differential calculus [Mor94, Mor06] to derive a sharp stationarity condition, where Robinson's strong regularity condition [Rob80] serves as a proper constraint qualification. We further develop a projected-gradient-type algorithm for computing a Clarke-type stationary point, which is slightly weaker than the Mordukhovich-type stationary point. Our numerical experiments will demonstrate applications in point spread function calibration and multiframe blind deconvolution.

Chapter 4 of the thesis is motivated from modeling a video clip, i.e. a sequence of image frames, by decomposing it into two "sparse" components of different natures. Once each frame of the image sequence is stacked as a single column of a matrix, we essentially speak of a matrix decomposition problem. More specifically, we aim to decompose, up to some small fitting error, the given data matrix (encoding original video contents) into a *low-rank matrix* sparse in singular values and a *sparse matrix* sparse in matrix entries. In the context of a surveillance video, a stationary background is typically modeled by the low-rank matrix, while moving objects are extracted by the sparse matrix. We remark that such a low-rank plus sparse matrix decomposition, in a more general context, is referred to as *robust principal component pursuit* (RPCP) in Candés et al [CLMW11], as RPCP robustifies the classical principle component analysis by taking into account extreme outliers with respect to the principle components. Concerning the numerical solution for RPCP, instead of invoking a convex-relaxation model based on the numerical norm and the ℓ^1-norm as most popular approaches in the literature do, in chapter 4 we consider a least-squares formulation subject to rank and cardinality constraints. An alternating minimization scheme is then employed to solve the resulting nonconvex constrained minimization problem. In particular, the low-rank matrix subproblem is resolved by a tailored Riemannian optimization technique [AMS08], which avoids singular value decompositions in full dimension. From the perspective of an inexact Riemannian Newton method, we establish a q-linear convergence theory for the overall alternating minimization scheme. Finally, we demonstrate numerical evidences that our newly proposed method compares favorably with the convex-relaxation based approach.

As the main body of the thesis, chapters 2–4 are all structured in a similar way. Each chapter begins with an introduction section, which describes the background of the problem under consideration and reviews the existing literature. Then a preliminary section provides a connection between the upfront research on the corresponding subject and the relevant mathematical tools at a more fundamental level. This is followed by the original research by the author of the present thesis that consists of a complete presentation of analyses, algorithms, and numerics. Most findings within this thesis have been published in academic journals from the corresponding fields; see [HW13, HW14a, HW14b, HW15b, HW15a]. Finally, chapter 5 concludes the thesis with a brief summary and an outlook on the future work. It should be noted that the notations and symbols used in the thesis are self-consistent within each individual chapter, which are typically clarified at the beginning of the presentations. The bibliography at the end of the thesis is ordered alphabetically according to the citation labels.

Chapter 2

Nonconvex TV^q-models in image restoration: analysis, algorithm, numerics, and generalizations

2.1 Introduction

In many applications of signal and image recovery one is interested in obtaining solutions with the sparsest or smallest support set, either of the signal directly or of a related quantity of interest (such as the gradient of an image for instance), from a limited number of measurements. This topic is at the core of *compressed sensing* (see, e.g., [CT06, DL92, DS89, DDFG10] and the references therein) or *basis pursuit* (see, e.g., [CDS98]) and has sparked significant research activities in the recent past. Mathematically, finding the smallest support set of a signal or an image requires to minimize the ℓ^0-norm, i.e. the number of nonzero entries in the solution vector or the related quantity of interest, subject to a constraint reflecting data fidelity. This problem is of combinatorial nature and it is well-known that it is essentially NP-hard [Nat95]. Thus, for practical purposes the ℓ^0-norm minimization problem is usually replaced by a convex relaxation leading to the minimization of the ℓ^1-norm which can be solved efficiently; see the discussion in [DDFG10] and, for instance, [TW10] and the references therein for further algorithmic developments.

In image processing one typically aims at recovering an image from noisy data while still keeping edges in the image. The latter requirement is responsible for the tremendous success of total variation based image restoration [ROF92]. In connection with the sparsity requirement alluded to above, this implies to compute a restoration result with gradient-sparsity, i.e. a piecewise constant image with a small number of patches. Hence, rather than minimizing the support of the image directly, one is interested in minimizing the support of the gradient of the recovered image. In the context of the convex relaxation mentioned above this amounts to minimizing the ℓ^1-norm of the gradient of the image subject to data fidelity; see, e.g., [CGM99, HK04, Nes05, HS06, GO09, TW10, BBC11] and the references therein for associated solution algorithms.

There is evidence [Cha07b, NNZC08] that replacing the ℓ^1-norm by the nonconvex and nondifferentiable function $\|v\|_{\ell^q}^q = \sum_i |v_i|^q$ with $q \in (0,1)$, which for the ease of reference we refer to as ℓ^q-norm in what follows, promotes gradient-sparsity even better. Moreover, the ℓ^q-norm allows possibly a smaller number of measurements than the ℓ^1-norm in compressed sensing. In [Nik02] (see also the more recent paper [NNZC08]) it was shown that nonconvex regularization terms in total variation based image restoration yield even better edge preservation when compared to the convex ℓ^1-type regularization. Moreover, it appears that the ℓ^q-norm regularization is also more robust with respect to noise.

Nonconvex and nonsmooth regularization in image restoration (and more generally in inverse problems) poses significant challenges with respect to both, the existence of solutions of associated minimization problems and, in particular, the development of efficient (i.e. locally more than the linearly convergent) solution algorithms. Linearly convergent gradient projection type methods for compressed sensing problems minimizing the ℓ^q-norm can be found in [Cha07b]. In [CY08] the latter solver was replaced by a regularized iteratively reweighted least squares (IRLS) technique. Based on [GO09], Chartrand extends in [Cha09] the Bregman iteration which relies on a variable splitting approach combined with a q-shrinkage operation to ℓ^q-norm minimization. The resulting method typically has a linear convergence behavior. In [DDFG10], the iteratively reweighted least squares solver for compressed sensing with the ℓ^q-norm is shown to converge locally superlinearly. The result depends on a q-null-space-condition, the sparsity of the solution and a locality requirement of the initial guess. A different perspective was taken in [NNZC08] where, under certain conditions, more general nonconvex regularization functionals are considered. Concerning the solver development, a technique based on interior point method is proposed. The authors make the interesting observation that, under the stated conditions, the nonsmooth and nonconvex regularization functional may be decomposed as the sum of a nonconvex but smooth part plus a convex and nonsmooth part. Increasing the variable space and rewriting the problem then yields the minimization of a nonconvex and smooth function subject to linear or affine equality constraints and nonnegativity constraints, which is equivalent to the original problem. The reformulated problem may now be tackled by interior point methods [Wri97], which were very recently shown to compute a local minimizer in compressed sensing in polynomial time [GJY11]. Clearly, the increase of the variable space and the computational effort implied by the interior point methods might be considered as disadvantages. In the follow-up work [NNT10] the interior point solver is replaced by variable splitting techniques resulting in alternating minimization methods which converge linearly. Unfortunately, the conditions required for the success of the algorithms proposed in [NNZC08] and [NNT10] rule out the ℓ^q-norm minimization and also the modified version of this problem considered in this chapter. We also mention the development of a smoothing nonlinear conjugate gradient solver in [CZ10] which is based on [NNZC08].

In this chapter we are interested in expanding the scope of solvers for ℓ^q-norm-based regularization of the gradient of the image to be recovered (we refer to this regularization as the TV^q-regularization as it combines the edge preservation property of total variation regularization with the sparsity-promoting ℓ^q-norm). In particular we are interested in locally superlinearly convergent methods which are robust with respect to noise. In order to achieve this, our proposed method considers a Huber-type regularization of the non-Lipschitz ℓ^q-norm and combines a reweighting technique for handling the nonconvexity with primal-dual semismooth Newton methods for image restoration [CGM99, HK04, HS06], which exhibit a fast (local) convergence towards a stationary point. For stabilizing the Newton solver in the presence of indefiniteness due to the involved nonconvexity, a specific regularization scheme is applied which modifies the (generalized) Hessian of the underlying variational problem based on a trust-region technique [CGT00]. The latter technique has the advantage of allowing a transition of the modified (generalized) Hessian to the true Hessian as the solution is approached and, thus, enabling the local superlinear convergence properties of the underlying Newton iteration. We point out that in contrast to the IRLS solver of [DDFG10] we guarantee global convergence. Moreover, local superlinear convergence is established without requiring conditions like the q-null-space property or sparsity conditions concerning the solution.

The rest of the chapter is organized as follows. Section 2.2 consists of preliminaries of some classical theories on the total-variation model as well as relevant numerical methods for nonconvex and nonsmooth minimizations. In section 2.3, we introduce our TV^q-model problem and discuss its regularization by a Huber-type function. The primal-dual Newton solver is the subject of section 2.4. In this core section of the present chapter, we introduce the stabilization of Newton's method (which we call R-regularization) together with the associated trust-region scheme for deciding on the amount of R-regularization required. Furthermore, the overall algorithm is defined and its global as well as local superlinear convergence is established. Section 2.5 is devoted to numerical tests showing the efficiency of our new method. A smoothing scheme to trace the original nonsmooth TV^q problem via a sequence of smoothed problems is provided in section 2.6. In section 2.7, we address the function space setting of the underlying variational problem and discuss the associated difficulties including a warning example. Finally, generalization of our TV^q-models to a more general class of variational models with concave priors is conducted in section 2.8.

2.2 Preliminaries

A systematic approach to investigating an optimization problem roughly consists of three steps. First, existence of optimal solutions needs to be justified, ideally in a properly chosen infinite dimensional space. In this respect, direct methods in the calculus of variations are typically the ways to follow. Once it exists, characterization of an optimal solution, often known as the (necessary) optimality condition, becomes interesting since this helps us qualify or disqualify certain candidate solutions among others. The final step is to devise a numerical scheme (often in a discrete setting) for computing an optimal solution in the sense that it is globally optimal for the underlying optimization problem or at least satisfies the derived optimality condition. Since such a numerical scheme is often iterative in nature, its convergence properties need to be carefully analyzed.

In this preliminary section, we recap some classical theories, under the context of the present chapter, in terms of the three aforementioned aspects. Section 2.2.1 concerns the existence of solutions and the optimality condition of the classical total-variation (TV) model in infinite dimensions, as the (convex) TV-model is the precursor of the nonconvex TV^q-models in the main body of this chapter. This is followed by two algorithmic subsections under the finite dimensional settings. A locally superlinearly convergent semismooth Newton method is introduced in section 2.2.2. In addition, two globalization strategies for nonconvex minimizations, namely the line search method and the trust-region method, are presented in section 2.2.3.

2.2.1 Functional analytic aspects of total-variation models

In section 2.2.1, we consider the following TV-model, see [CK97, HK04], in a continuous setting:

$$\min\ \alpha \int_\Omega |Du| + \frac{\widetilde{\mu}}{2}\int_\Omega |u|^2 dx + \frac{1}{2}\int_\Omega |Ku-z|^2 dx, \quad \text{over } u \in \mathrm{BV}(\Omega). \tag{2.2.1}$$

Here Ω denotes a domain in $\mathbb{R}^2$ which is bounded, simply connected, and has a Lipschitz boundary $\partial\Omega$. The observed image z is a square-integrable function on Ω, i.e. $z \in L^2(\Omega)$, and K is a known continuous linear map from $L^2(\Omega)$ to itself, i.e. $K \in \mathcal{L}(L^2(\Omega))$. The parameters $\alpha > 0$, $\widetilde{\mu} \geq 0$ are chosen by the user. We denote by dx the Lebesgue measure in $\mathbb{R}^2$, and without further specification the symbol "a.e." means almost everywhere with respect to this measure. We use the notation $K^\top$ for the adjoint of K. Besides, $\mathrm{BV}(\Omega)$ denotes the space of all functions of bounded variation (BV) on Ω with the associated BV-seminorm $\int_\Omega |Du|$; see (2.2.3) and (2.2.2) below for the corresponding definitions. For introductions to general theory on BV functions, we refer to [Giu84, ABM06]. The BV space is more appropriate in digital image modeling than Sobolev spaces, since a BV function admits discontinuities (often edges in an image) while excludes extensive oscillations (often noises in an image).

The rest of section 2.2.1 is organized as follows. We first define the BV space and prove the existence of solutions for (2.2.1). This is followed by the derivation of the optimality condition for (2.2.1) via the Fenchel duality. Finally, we introduce an approximation of the model (2.2.1) in a Hilbert space.

Existence of solutions in the BV space

Define the BV-seminorm of $u \in L^1(\Omega)$ by

$$\int_\Omega |Du| := \sup\left\{\int_\Omega u \,\mathrm{div} v\, dx : v \in C_c^1(\Omega;\mathbb{R}^2),\ |v(x)| \leq 1 \text{ a.e. in } \Omega\right\}. \tag{2.2.2}$$

Here $C_c^1(\Omega;\mathbb{R}^2)$ refers to the set of all continuously differentiable $\mathbb{R}^2$-valued functions on Ω with compact supports. The space $\mathrm{BV}(\Omega)$ is defined by

$$\mathrm{BV}(\Omega) := \left\{ u \in L^1(\Omega) : \int_\Omega |Du| < \infty \right\}. \tag{2.2.3}$$

Note that $\mathrm{BV}(\Omega)$ is a Banach space endowed with the norm $\|\cdot\|_{\mathrm{BV}(\Omega)}$ defined by $\|u\|_{\mathrm{BV}(\Omega)} := \|u\|_{L^1(\Omega)} + \int_\Omega |Du|$; see, e.g., Theorem 10.1.1 in [ABM06]. As a remark, the distributional derivative Du should be understood as a $\mathbb{R}^2$-valued Borel measure; see, e.g., Definition 10.1.1 in [ABM06].

Some important properties of the BV space are stated in the following two lemmas.

Lemma 2.2.1 (Weak lower semicontinuity in $L^p(\Omega)$, $1 \le p < \infty$)**.** *The BV-seminorm $u \mapsto \int_\Omega |Du|$ is weakly lower semicontinuous in $L^p(\Omega)$ for $1 \le p < \infty$, i.e. for any $u \in L^p(\Omega)$ we have*

$$\int_\Omega |Du| \le \liminf \left\{ \int_\Omega |Du^k| : u^k \rightharpoonup u \text{ in } L^p(\Omega) \right\}.$$

Proof. See Theorem 2.3 in [AV94]. □

Lemma 2.2.2 (Embedding of $\mathrm{BV}(\Omega)$ into $L^p(\Omega)$, $1 \le p \le 2$)**.** *The space $\mathrm{BV}(\Omega)$ is continuously embedded into $L^p(\Omega)$ for $1 \le p \le 2$, i.e. there exists a constant C_p depending on Ω, p only such that the following inequality holds for all $u \in \mathrm{BV}(\Omega)$:*

$$\left(\int_\Omega |u|^p dx \right)^{1/p} \le C_p \|u\|_{\mathrm{BV}(\Omega)}.$$

Furthermore, if $1 \le p < 2$, the embedding is even compact, i.e. every bounded sequence in $\mathrm{BV}(\Omega)$ has a (strongly) convergent subsequence in $L^p(\Omega)$.

Proof. This result is a special case of Theorems 10.1.3 and 10.1.4 in [ABM06] for $\Omega \subset \mathbb{R}^2$. □

Now we show the existence of a solution for the TV-model (2.2.1).

Theorem 2.2.3 (Existence of solution)**.** *Assume that $\widetilde{\mu} > 0$ or $K^\top K$ is nonsingular. Then (2.2.1) admits a unique global minimizer.*

Proof. Our proof closely follows Theorem 2.1 in [CK97].

Provided that a global minimizer exists, its uniqueness follows immediately from the strict convexity under our assumptions. To show the existence, let $\{u^k\} \subset \mathrm{BV}(\Omega)$ be an infimizing sequence for (2.2.1).

We claim that $\{u^k\}$ is uniformly bounded in $\mathrm{BV}(\Omega)$. If this is not true, we have, possibly along a subsequence, that $\|u^k\|_{L^1(\Omega)} \to \infty$ or $\int_\Omega |Du^k| \to \infty$. In both cases, the objective in (2.2.1) tends to infinity, contradicting the assumption that $\{u^k\}$ is an infimizing sequence.

Thanks to the embedding in Lemma 2.2.2, there exists a subsequence $\{u^k\}$ which converges to some u^* strongly in $L^1(\Omega)$ and weakly in $L^2(\Omega)$. Note that $u^* \in \mathrm{BV}(\Omega)$ can be checked from

Lemma 2.2.1. Moreover, the mapping $u \mapsto \alpha \int_\Omega |Du| + \frac{\widetilde{\mu}}{2} \int_\Omega |u|^2 dx$ is convex and (strongly) continuous in $L^2(\Omega)$, and therefore also weakly lower semicontinuous in $L^2(\Omega)$. Together with Lemma 2.2.1, the objective in (2.2.1) is weakly lower semicontinuous in $L^2(\Omega)$, and we conclude that u^* is a global minimizer. $\square$

Optimality condition via Fenchel duality

Our next concern is to derive an optimality condition for the TV-model (2.2.1) based on the Fenchel duality theorem. For this purpose, we first introduce some notions from convex analysis (in infinite dimensions), for which further information can be found in [ET99].

Let X be a Banach space with its topological dual X^*, and $\langle \cdot, \cdot \rangle_{X,X^*}$ be a duality pairing over $X \times X^*$. The function $\Theta^* : X^* \to \mathbb{R} \cup \{\infty\}$ is called the convex conjugate of a convex function $\Theta : X \to \mathbb{R} \cup \{\infty\}$ if we have for all $\widehat{u} \in X^*$ that

$$\Theta^*(\widehat{u}) = \sup_{u \in X} \left\{ \langle u, \widehat{u} \rangle_{X,X^*} - \Theta(u) \right\}.$$

In convex analysis, the subdifferential of Θ at u, denoted by $\partial\Theta(u)$, is defined by

$$\partial\Theta(u) := \{ \widehat{u} \in X^* : \Theta(v) \geq \Theta(u) + \langle v - u, \widehat{u} \rangle_{X,X^*} \ \forall v \in X \}$$

if $\Theta(u) < \infty$, and $\partial\Theta(u) := \emptyset$ if $\Theta(u) = \infty$.

We state the Fenchel duality theorem in the following. For convenience of our later utilization, we set p as the primal variable and u as the dual variable in our formulation.

Theorem 2.2.4 (Fenchel duality)**.** *Let X and Y be two Banach spaces with topological duals X^* and Y^*, respectively, and $\Lambda \in \mathcal{L}(X;Y)$. Assume that $\Theta : X \to \mathbb{R} \cup \{\infty\}$ and $\Psi : Y \to \mathbb{R} \cup \{\infty\}$ are convex lower semicontinuous functionals and there exists $p_0 \in X$ such that $\Theta(p_0) < \infty$, $\Psi(\Lambda p_0) < \infty$, and Ψ is continuous at Λp_0. Then we have*

$$\inf_{p \in X} \Theta(p) + \Psi(\Lambda p) = \sup_{u \in Y^*} -\Theta^*(-\Lambda^\top u) - \Psi^*(u), \tag{2.2.4}$$

where $\Theta^* : X^* \to \mathbb{R} \cup \{\infty\}$ and $\Psi^* : Y^* \to \mathbb{R} \cup \{\infty\}$ are the convex conjugates of Θ and Ψ, respectively. Moreover, the optimal solutions $p^* \in X$ and $u^* \in Y^*$ in (2.2.4) are characterized by the following optimality conditions:

$$\begin{cases} -\Lambda^\top u^* \in \partial\Theta(p^*), \\ \quad \Lambda p^* \in \partial\Psi^*(u^*). \end{cases}$$

Proof. See pp. 60 in [ET99]. $\square$

In the following, we will associate (2.2.1) with the following predual problem:

$$\begin{aligned} &\min \ \frac{1}{2} \left\langle v + K^\top z, (\widetilde{\mu} I + K^\top K)^{-1}(v + K^\top z) \right\rangle_{L^2(\Omega)} \\ &\text{s.t. } v \in H_0^{\text{div}}(\Omega), \ |v(x)| \leq 1 \text{ a.e. in } \Omega. \end{aligned} \tag{2.2.5}$$

Let $\mathcal{H}^1$ be the one-dimensional (outer) Hausdorff measure; see, e.g., pp. 110 in [ABM06]. The space $H_0^{\mathrm{div}}(\Omega)$ is defined by

$$H_0^{\mathrm{div}}(\Omega) := \left\{p \in L^2(\Omega;\mathbb{R}^2) : \mathrm{div} p \in L^2(\Omega),\ p(x)\cdot\nu(x) = 0 \text{ a.e. with respect to } \mathcal{H}^1 \text{ on } \partial\Omega\right\},$$

where ν is the outer normal vector on $\partial\Omega$. To utilize Theorem 2.2.4, we set

$$\begin{aligned}
X &:= H_0^{\mathrm{div}}(\Omega),\\
Y = Y^* &:= L^2(\Omega),\\
\Lambda &:= \alpha\,\mathrm{div},\\
\Sigma_p &:= \{p \in H_0^{\mathrm{div}}(\Omega) : |p(x)| \le 1 \text{ a.e. in } \Omega\},\\
\Theta(p) &:= \begin{cases} 0 & \text{if } p \in \Sigma_p,\\ \infty & \text{otherwise,}\end{cases}\\
\Psi(v) &:= \frac{1}{2}\left\langle v + K^\top z, (\widetilde{\mu} I + K^\top K)^{-1}(v + K^\top z)\right\rangle_{L^2(\Omega)}.
\end{aligned}$$

The convex conjugates of Θ and Ψ can be readily calculated as

$$\begin{aligned}
\Theta^*(w) &= \sup\left\{\langle p, w\rangle_{H_0^{\mathrm{div}}(\Omega),H_0^{\mathrm{div}}(\Omega)^*} : p \in \Sigma_p\right\},\\
\Psi^*(u) &= \frac{\widetilde{\mu}}{2}\int_\Omega |u|^2 dx + \frac{1}{2}\int_\Omega |Ku - z|^2 dx.
\end{aligned}$$

One may consider $\mathrm{div} \in \mathcal{L}(H_0^{\mathrm{div}}(\Omega); L^2(\Omega))$ and $\nabla \in \mathcal{L}(L^2(\Omega); H_0^{\mathrm{div}}(\Omega)^*)$ such that $\nabla := -\mathrm{div}^\top$. Since $\{p \in C_c^1(\Omega;\mathbb{R}^2) : |p(x)| \le 1 \text{ a.e. in } \Omega\}$, i.e. the feasible set for the supremum in (2.2.2), is dense in Σ_p under the (strong) $H_0^{\mathrm{div}}(\Omega)$-topology, see Theorem 2 in [HR15], we have

$$\Theta^*(\alpha\nabla u) = \alpha\int_\Omega |Du|$$

for any $u \in L^2(\Omega)$. In fact, $\Theta^*(\alpha\nabla u) + \Psi^*(u) < \infty$ if and only if $u \in \mathrm{BV}(\Omega)$.

Thus, based on the Fenchel duality, we arrive at the optimality condition stated in the following theorem.

Theorem 2.2.5 (Optimality condition)**.** *The Fenchel dual of (2.2.5) is given by (2.2.1). Moreover, the optimal solutions u^* and p^* for (2.2.1) and (2.2.5), respectively, satisfy the following conditions:*

$$\begin{cases} (\widetilde{\mu} I + K^\top K)u^* - \alpha\,\mathrm{div} p^* = K^\top z & \text{in } L^2(\Omega),\\ \langle p - p^*, \nabla u^*\rangle_{H_0^{\mathrm{div}}(\Omega),H_0^{\mathrm{div}}(\Omega)^*} \le 0 & \forall p \in \Sigma_p.\end{cases}$$

A Hilbert-space approach

Now we consider the approximation of the TV-model (2.2.1) (with $\widetilde{\mu} = 0$) in the Hilbert space $H_0^1(\Omega)$, see related work in [IK99, HS06]. Our variational model in $H_0^1(\Omega)$ is formulated as follows:

$$\min\ \alpha\int_\Omega |\nabla u| dx + \frac{\mu}{2}\int_\Omega |\nabla u|^2 dx + \frac{1}{2}\int_\Omega |Ku - z|^2 dx, \quad \text{over } u \in H_0^1(\Omega). \tag{2.2.6}$$

Here ∇u is the distributional derivative of u. We shall denote the topological dual of $H_0^1(\Omega)$ by $H^{-1}(\Omega)$ and the Laplacian by $\Delta := \operatorname{div} \circ \nabla \in \mathcal{L}(H_0^1(\Omega); H^{-1}(\Omega))$. Different from (2.2.1), there is an additional H^1-term with the leading coefficient $0 \leq \mu \ll \alpha$. The space $H_0^1(\Omega)$ consists of all functions $u \in L^2(\Omega)$ such that

$$\|u\|_{H_0^1(\Omega)} := \left(\int_\Omega |\nabla u|^2 dx \right)^{1/2} < \infty,$$

and u has zero trace over $\partial\Omega$. Note that $\int_\Omega |\nabla u| dx = \int_\Omega |Du|$ for $u \in H_0^1(\Omega)$.

Under the assumption that $\mu > 0$ or $K^\top K$ is nonsingular, existence and uniqueness of the solution for (2.2.6) can be verified using standard arguments from the direct methods of the calculus of variations. Besides, the optimality condition for (2.2.6) is again a consequence of the Fenchel duality in Theorem 2.2.4.

Theorem 2.2.6 (Optimality condition)**.** *The optimal solution u^* for (2.2.6) satisfies the following conditions for some $p^* \in L^2(\Omega; \mathbb{R}^2)$:*

$$\begin{cases} -\mu\Delta u^* + K^\top K u^* - \alpha \operatorname{div} p^* = K^\top z \quad \text{in } H^{-1}(\Omega), \\ \left.\begin{array}{r} |(\nabla u^*)(x)| p^*(x) = (\nabla u^*)(x) \quad \text{if } |p^*(x)| = 1 \\ (\nabla u^*)(x) = 0 \quad \text{if } |p^*(x)| < 1 \end{array}\right\} \quad \text{for } p^* \in L^2(\Omega; \mathbb{R}^2). \end{cases}$$

Proof. See Theorem 2.1 in [HS06]. □

It is justified in the following theorem that the H^1-model (2.2.6) is indeed a faithful approximation to the original TV-model (2.2.1).

Theorem 2.2.7 (Consistency)**.** *Without loss of generality, assume that $\widetilde{\mu} = 0$ and K is the identity. Let $\{\mu^k\}$ be a sequence of positive scalars such that $\mu^k \to 0^+$. For each μ^k, let $u^k \in H_0^1(\Omega)$ be the optimal solution of (2.2.6) with $\mu := \mu^k$. Then there exists a weak accumulation point u^* of $\{u^k\}$ in $L^2(\Omega)$ such that u^* is the optimal solution for (2.2.1).*

Proof. Our proof is based on Remark 2.1 in [IK99].

First, note that $\{u^k\}$ is uniformly bounded in $\mathrm{BV}(\Omega)$. Hence, by Lemma 2.2.2, there exists a subsequence, again denoted by $\{u^k\}$, such that u^k converges to u^* strongly in $L^1(\Omega)$ and weakly in $L^2(\Omega)$. In addition, for each k we have

$$\alpha \int_\Omega |\nabla u^k| dx + \frac{\mu^k}{2} \int_\Omega |\nabla u^k|^2 dx + \frac{1}{2} \int_\Omega |u^k - z|^2 dx \leq \alpha \int_\Omega |\nabla \widetilde{u}| dx + \frac{\mu^k}{2} \int_\Omega |\nabla \widetilde{u}|^2 dx + \frac{1}{2} \int_\Omega |\widetilde{u} - z|^2 dx$$

for all $\widetilde{u} \in H_0^1(\Omega)$. Fixing $\widetilde{u}$ and taking the limit inferior with respect to k on both sides of the above inequality, we have

$$\alpha \int_\Omega |Du^*| + \frac{1}{2} \int_\Omega |u^* - z|^2 dx \leq \alpha \int_\Omega |\nabla \widetilde{u}| dx + \frac{1}{2} \int_\Omega |\widetilde{u} - z|^2 dx, \tag{2.2.7}$$

due to the weak lower semicontinuity in Lemma 2.2.1. Note that (2.2.7) holds true for an arbitrary $\widetilde{u} \in H_0^1(\Omega)$. Furthermore, given an arbitrary $u \in \mathrm{BV}(\Omega)$, there exists $\{\widetilde{u}^l\}$ in $H_0^1(\Omega)$

such that $\widetilde{u}^l$ converges to u (strongly) in $L^2(\Omega)$ and $\lim_{l\to\infty}\int_\Omega |D\widetilde{u}^l| = \int_\Omega |Du|$; see Theorem 1.17 in [Giu84]. Thus, (2.2.7) also holds true for any $\widetilde{u} \in \mathrm{BV}(\Omega)$, i.e. u^* is optimal for (2.2.1). □

As a remark, we mention that the H^1-model (2.2.6) is convenient for numerical purposes. In [HS06], up to a Tikhonov regularization on the dual problem for (2.2.6), an efficient semismooth Newton method is developed to compute the optimal solution. We also remark that the H^1-model (2.2.6) is the precursor of the TV^q-model in (2.3.1).

2.2.2 Semismooth Newton method

In section 2.2.2, we introduce the semismooth Newton method for (iteratively) solving the nonlinear equation

$$F(u) = 0, \tag{2.2.8}$$

where the associated operator $F : \mathbb{R}^n \to \mathbb{R}^m$ is merely locally Lipschitz (rather than continuously differentiable). In this regard, the semismooth Newton method generalizes the classical Newton's method. Under the finite dimensional settings, our following presentation on the semismooth Newton method is based on [QS93, IK08].

Based on Rademacher's theorem, which asserts that in finite dimensions any locally Lipschitz function is differentiable almost everywhere, we introduce two notions of generalized derivatives. The B(ouligand)-subdifferential of F is defined by

$$\partial_B F(u) := \left\{ \lim_{k\to\infty} DF(u^k) : u^k \to u,\ F \text{ is differentiable at } u^k \text{ for each } k \right\},$$

and the (Clarke) subdifferential $\partial F(u)$ is defined as the convex hull of $\partial_B F(u)$. Then the semismooth Newton method can be described by the iteration formula below for $k = 0, 1, 2, ...$:

$$u^{k+1} := u^k - (V^k(u^k))^{-1} F(u^k), \quad \text{where } V^k(u^k) \in \partial_B F(u^k) \text{ is nonsingular.} \tag{2.2.9}$$

It will be shown in Theorem 2.2.12 that the semismooth property of F leads to the local superlinear convergence of the iterative scheme (2.2.9).

Definition 2.2.8 (Directionally differentiable function). *A function $F : \mathbb{R}^n \to \mathbb{R}^m$ is directionally differentiable at $u \in \mathbb{R}^n$ along $d \in \mathbb{R}^n$ if*

$$DF(u; d) := \lim_{t\to 0^+} \frac{F(u + td) - F(u)}{t}$$

exists. We say F is directionally differentiable at u if $DF(u; d)$ exists for any $d \in \mathbb{R}^n$.

Definition 2.2.9 (Semismooth function). *A function $F : \mathbb{R}^n \to \mathbb{R}^m$ is said to be semismooth at u if F is locally Lipschitz at u and the following limit exists for all $d \in \mathbb{R}^n$:*

$$\lim_{\substack{V(u+t\widetilde{d}) \in \partial F(u+t\widetilde{d}),\\ \widetilde{d}\to d,\ t\to 0^+}} V(u + t\widetilde{d})\widetilde{d}.$$

Lemma 2.2.10. *Assume that $F : \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz at $u \in \mathbb{R}^n$. Then the following statements are equivalent:*

1. *F is semismooth at u.*

2. *F is directionally differentiable at u, and it holds for each $V(u+d) \in \partial F(u+d)$ that*

$$\|V(u+d)d - DF(u;d)\| = o(\|d\|), \quad \text{as } d \to 0. \tag{2.2.10}$$

Proof. See Theorem 2.3 in [QS93]. □

Lemma 2.2.11. *Assume that $F : \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz at $u \in \mathbb{R}^n$ and all $V(u) \in \partial_B F(u)$ are nonsingular. Then there exist a neighborhood U_u of u and a positive constant C_F such that $\|V(\widetilde{u})^{-1}\| \leq C_F$ for all $\widetilde{u} \in U_u$ and $V(\widetilde{u}) \in \partial_B F(\widetilde{u})$.*

Proof. The proof is analogous to that for Proposition 3.1 in [QS93], though that proposition requires a stronger assumption, i.e. $V(u) \in \partial F(u)$, than the present lemma. □

Theorem 2.2.12 (Local superlinear convergence)**.** *Let $u^* \in \mathbb{R}^n$ be a solution of (2.2.8). Further assume that F is semismooth at u^* and all $V(u^*) \in \partial_B F(u^*)$ are nonsingular. Let the iterating sequence $\{u^k\}$ be generated by formula (2.2.9) starting from an initial guess u^0 sufficiently close to u^*. Then the sequence $\{u^k\}$ converges superlinearly to u^*, i.e. $\lim_{k\to\infty} u^k = u^*$ and*

$$\lim_{k\to\infty} \frac{\|u^{k+1} - u^*\|}{\|u^k - u^*\|} = 0.$$

Proof. Our proof is analogous to Theorem 3.2 in [QS93].

In view of Lemma 2.2.11, formula (2.2.9) is well defined for each k and, moreover, $\{(V^k(u^k))^{-1}\}$ is uniformly bounded. Then we have

$$\begin{aligned}
\|u^{k+1} - u^*\| &= \|u^k - u^* - (V^k(u^k))^{-1} F(u^k)\| \\
&\leq \|(V^k(u^k))^{-1}(F(u^k) - F(u^*) - DF(u^*; u^k - u^*))\| \\
&\quad + \|(V^k(u^k))^{-1}(V^k(u^k)(u^k - u^*) - DF(u^*; u^k - u^*))\| \\
&= o(\|u^k - u^*\|), \quad \text{as } \|u^k - u^*\| \to 0.
\end{aligned}$$

The last equality above follows from Defintion 2.2.8, Lemma 2.2.10, and the uniform boundedness of $\{(V^k(u^k))^{-1}\}$. Since u^0 is assumed to be sufficiently close to u^*, the proof is complete. □

We conclude section 2.2.2 by noting that an analogous semismooth Newton method can be posed in infinite dimensional spaces, but the theoretical framework differs from its counterpart in finite dimensions due to lack of Rademacher's theorem in infinite dimensions; see [CNQ00, HIK03, Ulb03].

2.2.3 Globalization strategies for nonconvex minimizations

It is known that the semismooth Newton method in section 2.2.2 typically enjoys fast local convergence for obtaining a stationary point of the unconstrained minimization

$$\min_{u \in \mathbb{R}^n} f(u).$$

Throughout section 2.2.3, we assume that the objective $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function bounded from below and its derivative ∇f is locally Lipchitz. However, global convergence is not guaranteed for the Newton-type method from an arbitrary initial guess provided that the objective f is nonconvex. This calls for so-called globalization on iterative algorithms. Here we present two classical globalization strategies, namely the line search method and the trust-region method. By stipulating proper conditions on the update step, global convergence can be proven for both methods. Moreover, we remark that a state-of-the-art optimization algorithm should always satisfy a sufficient condition for global convergence, while asymptotically function like a Newton-type method in order to attain local superlinear convergence. The materials in section 2.2.3 are collected from standard optimization textbooks [DS96] and [NW06], where more implementation details as well as some historical perspectives can be traced.

Line search method

The iteration formula for a line search method is given by

$$u^{k+1} := u^k + a^k d^k. \tag{2.2.11}$$

Here u^k is the current iterate where $g^k := \nabla f(u^k) \neq 0$, d^k is a descent direction, i.e. $(g^k)^\top d^k < 0$, and $a^k > 0$ is the step size along d^k. After fixing the search direction d^k, the line search method selects a^k such that the following conditions are satisfied for some constants $0 < \tau_1 < \tau_2 < 1$:

$$f(u^{k+1}) \leq f(u^k) + \tau_1 a^k (g^k)^\top d^k, \tag{2.2.12}$$

$$\nabla f(u^{k+1})^\top d^k \geq \tau_2 (g^k)^\top d^k. \tag{2.2.13}$$

The inequality (2.2.12) alone is typically referred to as the Armijo condition. The line search method which fulfills both inequalities in (2.2.12)–(2.2.13) is called the Wolfe-Powell line search. In the following, we present the global convergence theory for the Wolfe-Powell line search.

Lemma 2.2.13. *There always exists an interval such that each a^k in this interval satisfies the Wolfe-Powell conditions (2.2.12)–(2.2.13).*

Proof. See Theorem 6.3.2 in [DS96]. □

Theorem 2.2.14 (Zoutendijk's theorem)**.** *For each $k \in \mathbb{N}$, let $a^k > 0$ fulfill the Wolfe-Powell conditions (2.2.12)–(2.2.13) and define $\cos\theta^k := -(g^k)^\top d^k/(\|g^k\|\|d^k\|)$. Then it follows that*

$$\sum_{k\in\mathbb{N}} \cos^2\theta^k \|g^k\|^2 < \infty.$$

Furthermore, provided that there exists a positive constant ϵ_c such that $\cos\theta^k \geq \epsilon_c$ for all $k \in \mathbb{N}$, we have the following global convergence:

$$\lim_{k\to\infty} \|\nabla f(u^k)\| = 0.$$

Proof. See Theorem 6.3.3 in [DS96]. □

Trust-region method

Different from the line search method, the trust-region method selects the update step (in terms of both step size and direction) by minimizing a local quadratic model subject to a trust-region constraint, i.e.

$$\begin{array}{ll} \min & h^k(d) := f(u^k) + (g^k)^\top d + \frac{1}{2} d^\top H^k d \\ \text{s.t.} & d \in \mathbb{R}^n, \ \|d\| \leq \sigma^k. \end{array} \tag{2.2.14}$$

Here $\sigma^k > 0$ is called the trust-region radius. The matrix H^k can be the Hessian $\nabla^2 f(u^k)$ or its approximation provided that f is twice continuously differentiable at u^k. Nevertheless, the global convergence of the trust-region method does not require any specific structural information on H^k. We specify the trust-region method in our discussion as follows.

Algorithm 2.2.15 (Trust-region method)**.**
Fix $0 < \rho_1 \leq \rho_2 < 1$, $0 < \kappa_1 < 1 < \kappa_2$. Initialize $u^0 \in \mathbb{R}^n$, $\sigma^0 > 0$. Iterate with $k = 0, 1, 2, ...$:

1. Generate d^k as an approximate solution for (2.2.14) such that the following Cauchy-point-based model reduction criterion is satisfied for some constant $0 < C_h \leq 1$:

   $$h^k(0) - h^k(d^k) \geq C_h \|g^k\| \min\left(\sigma^k, \frac{\|g^k\|}{\|H^k\|}\right). \tag{2.2.15}$$

   Then set $u^{k+1} := u^k + d^k$.

2. Evaluate the ratio $\rho^k := \dfrac{f(u^k) - f(u^{k+1})}{h^k(0) - h^k(d^k)}$.

3. Update the trust-region radius according to ρ^k, i.e.

   $$\sigma^{k+1} := \begin{cases} \kappa_1 \sigma^k & \text{if } \rho^k < \rho_1, \\ \kappa_2 \sigma^k & \text{if } \rho^k > \rho_2, \\ \sigma^k & \text{otherwise.} \end{cases}$$

To obtain a qualified d^k in step 1 of Algorithm 2.2.15, one can employ, e.g., the dogleg method or the truncated conjugate gradient method. We refer to [NW06, CGT00] for more elaborate introductions on trust-region subproblem solvers. The global convergence of Algorithm 2.2.15 is asserted in the following theorem.

Theorem 2.2.16 (Global convergence). *Let $\{u^k\}$ be generated by Algorithm 2.2.15. Assume that for each k, (2.2.15) and the following two inequalities all hold true for some constants $0 < C_h \leq 1$, $C_H > 0$, $C_\sigma \geq 1$:*

$$\|H^k\| \leq C_H, \quad \|d^k\| \leq C_\sigma \sigma^k.$$

Then we have the following global convergence:

$$\liminf_{k\to\infty} \|\nabla f(u^k)\| = 0.$$

Proof. See Theorem 4.5 in [NW06]. □

2.3 TV^q-model and its Huberization

We consider the following variational problem

$$\min_{u\in\mathbb{R}^{|\Omega|}} f(u) := \sum_{(i,j)\in\Omega} \left(\frac{\mu}{2}|(\nabla u)_{ij}|^2 + \frac{\alpha}{q}|(\nabla u)_{ij}|^q + \frac{\lambda_{ij}}{2}|(Ku-z)_{ij}|^2 \right), \tag{2.3.1}$$

where Ω is a two-dimensional index set representing the image domain. By $|\Omega|$ we denote its cardinality. We have $\alpha > 0$, $0 < q < 1$, $0 < \mu \ll \alpha$ as the given model parameters. The matrix $K \in \mathbb{R}^{|\Omega|\times|\Omega|}$ is assumed to not annihilate a constant vector, e.g. K might be a blurring matrix. The vector $z \in \mathbb{R}^{|\Omega|}$ stands for the given noisy data, and $u \in \mathbb{R}^{|\Omega|}$ is the image to be restored. Despite the fact that we refer to $u \in \mathbb{R}^{|\Omega|}$ as a vector, we denote the elements of u by u_{ij} with $(i,j) \in \Omega$. This appears natural as the image domain is given as a two-dimensional array of pixels. Analogously one has to understand the action of the blurring operator (matrix) K. Notably we allow situations where the fidelity coefficient $\lambda \in \mathbb{R}^{|\Omega|}$ is possibly spatially dependent (see, e.g., [DHRC11a, DHRC11b]) such that $\lambda_{ij} > 0$ for all $(i,j) \in \Omega$ and $\sum_{(i,j)\in\Omega} \lambda_{ij} = |\Omega|$, though $\lambda_{ij} = 1$ for all $(i,j) \in \Omega$ is taken in the numerics. The discrete gradient operator ∇ is decomposed as $\nabla = (\nabla_x, \nabla_y)$ such that $(\nabla u)_{ij} = ((\nabla_x u)_{ij}, (\nabla_y u)_{ij})$, where $\nabla_x \in \mathbb{R}^{|\Omega|\times|\Omega|}$ is the discrete derivative in x-direction and $\nabla_y \in \mathbb{R}^{|\Omega|\times|\Omega|}$ is the discrete derivative in y-direction, respectively. The Euclidean norm of $(\nabla u)_{ij}$ in $\mathbb{R}^2$ is denoted by $|(\nabla u)_{ij}|$. For elements $p \in (\mathbb{R}^{|\Omega|})^2$, p_x denotes components corresponding to the x-direction in the above sense and p_y components belonging to the y-direction. The discrete Laplacian Δ is defined as $\Delta := -\nabla_x^\top \nabla_x - \nabla_y^\top \nabla_y$. The multiplication of vectors is understood in the pointwise sense, i.e. $(uv)_{ij} = u_{ij}v_{ij}$ for $u, v \in \mathbb{R}^{|\Omega|}$ and $(up)_{ij} = (u_{ij}(p_x)_{ij}, u_{ij}(p_y)_{ij})$ for $u \in \mathbb{R}^{|\Omega|}$, $p \in (\mathbb{R}^{|\Omega|})^2$. Similarly, for $u \in \mathbb{R}^{|\Omega|}$ and $q \in \mathbb{R}$, the power u^q is a vector in $\mathbb{R}^{|\Omega|}$ such that $(u^q)_{ij} = u_{ij}^q$. For $u, v \in \mathbb{R}^{|\Omega|}$ and $\gamma \in \mathbb{R}$, the max-operation is understood in a componentwise sense, i.e. $(\max(u,\gamma))_{ij} = \max(u_{ij},\gamma)$ and $(\max(u,v))_{ij} = \max(u_{ij}, v_{ij})$. A diagonal matrix with its diagonal elements given by the vector u is denoted by $\operatorname{diag}(u)$. The characteristic vector $\chi_\mathcal{A}$ of the set $\mathcal{A} \subset \Omega$ is defined as $(\chi_\mathcal{A})_{ij} = 1$ if $(i,j) \in \mathcal{A}$ and $(\chi_\mathcal{A})_{ij} = 0$ otherwise. The identity vector $e \in (\mathbb{R}^{|\Omega|})^2$ is defined as $e_{ij} = (1,1)$ for all $(i,j) \in \Omega$. We use $\|\cdot\|$ to denote the 2-norm for vectors in $\mathbb{R}^{|\Omega|}$

and the spectral norm for matrices in $\mathbb{R}^{|\Omega|\times|\Omega|}$. The symbols $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ represent the maximal eigenvalue and the minimal eigenvalue of a matrix, respectively. The constant C may take different values at different occasions.

We start our investigations of (2.3.1) by establishing the existence of a solution.

Theorem 2.3.1 (Existence of solution)**.** *Assume that* $\mu \geq 0,\ \alpha > 0,\ q > 0,\ \lambda_{ij} > 0$ *for all* $(i,j) \in \Omega$*, and*

$$\operatorname{Ker}\nabla \cap \operatorname{Ker} K = \{0\}. \tag{2.3.2}$$

Then there exists a global minimizer for the variational problem (2.3.1).

Proof. Since f is bounded from below, it suffices to show that f is coercive, i.e. $|f(u^k)| \to \infty$ whenever $\|u^k\| \to \infty$ for some sequence (u^k) in $\mathbb{R}^{|\Omega|}$. We prove this by contradiction. For this purpose, assume that $\|u^k\| \to \infty$ and that $f(u^k)$ is uniformly bounded. For each k, let $u^k = s^k v^k$ such that $s^k \geq 0$, $v^k \in \mathbb{R}^{|\Omega|}$, and $\|v^k\| = 1$. Then we have

$$\lim_{k\to\infty} \sum_{(i,j)\in\Omega} \Big(\alpha|(\nabla v^k)_{ij}|^q/q + \lambda_{ij}|(Kv^k)_{ij}|^2/2\Big) = 0,$$

due to the fact that the functions $s \mapsto |s|^q$ and $s \mapsto |s|^2$ are both coercive. By compactness, the sequence (v^k) has an accumulation point v^* with $\|v^*\| = 1$ such that $v^* \in \operatorname{Ker}\nabla \cap \operatorname{Ker}K$. This contradicts our hypothesis (2.3.2). □

In order to characterize an optimal solution u, we define the active set $\mathcal{A}(u) := \{(i,j) \in \Omega : |(\nabla u)_{ij}| \neq 0\}$ and the inactive set $\mathcal{I}(u) := \Omega\backslash\mathcal{A}(u)$. Due to the occurrence of the term involving q in (2.3.1) with $0 < q < 1$ (which we call the TV^q-term from now on), the objective f (which we refer to as the TV^q-model) is nondifferentiable on $\mathcal{I}(u)$. Therefore, the Euler-Lagrange equation for characterizing a stationary point is separately posed on $\mathcal{A}(u)$ and on $\mathcal{I}(u)$, i.e.

$$\begin{cases} -\mu\Delta u + K^\top\lambda(Ku - z) + \alpha\nabla^\top(|\nabla u|^{q-2}\nabla u) = 0, & \text{if } (i,j) \in \mathcal{A}(u), \\ \nabla u = 0, & \text{if } (i,j) \in \mathcal{I}(u). \end{cases} \tag{2.3.3}$$

Since the objective f is nonconvex, the solution to (2.3.3) is in general not unique.

In order to make the problem numerically tractable, we locally smooth the TV^q-term by a *Huber function* φ_γ defined by

$$\varphi_\gamma(s) := \begin{cases} \frac{1}{q}|s|^q - (\frac{1}{q} - \frac{1}{2})\gamma^q, & \text{if } |s| \geq \gamma, \\ \frac{1}{2}\gamma^{q-2}|s|^2, & \text{if } |s| \leq \gamma. \end{cases}$$

Correspondingly, the *Huberized* variational problem is written as

$$\min_{u\in\mathbb{R}^{|\Omega|}} f_\gamma(u) := \sum_{(i,j)\in\Omega} \left(\frac{\mu}{2}|(\nabla u)_{ij}|^2 + \alpha\varphi_\gamma(|(\nabla u)_{ij}|) + \frac{\lambda_{ij}}{2}|(Ku - z)_{ij}|^2\right). \tag{2.3.4}$$

Note that the Huberized functional f_γ is continuously differentiable, and the Euler-Lagrange equation associated with (2.3.4) is given by

$$\nabla f_\gamma(u) = -\mu\Delta u + K^\top \mathrm{diag}(\lambda)(Ku - z) + \alpha\nabla^\top \left(\max(|\nabla u|, \gamma)^{q-2}\nabla u\right) = 0. \tag{2.3.5}$$

The Huber function [Hub64], as a tool of local smoothing, has been previously applied and analyzed on convex nondifferentiable variational models in image processing; see, e.g., [Vog02, HS06, DHN09]. For different nonconvex models with either smoothing or continuation we refer to, e.g., [Nik99, CZ10]. Next we study the behavior of our Huberization of the nonconvex TV^q-model for vanishing Huber parameter, i.e. for $\gamma \to 0^+$.

Theorem 2.3.2 (Consistency of Huberization)**.** *Let the assumptions of Theorem 2.3.1 hold true. Further assume that* (u^k) *is a uniformly bounded sequence with each* u^k *a stationary point of the Huberized problem (2.3.4) satisfying (2.3.5). Then as* $\gamma^k \to 0^+$*, there exists a subsequence of* (u^k) *converging to some* $u^* \in \mathbb{R}^{|\Omega|}$*, which satisfies the original Euler-Lagrange equation (2.3.3).*

Proof. By compactness, there exists a subsequence of (u^k), say $(u^{k'})$, such that $(u^{k'})$ converges to some u^* as $k' \to \infty$. Next we show that u^* is a solution to (2.3.3). Since each $(u^{k'})$ satisfies the Huberized Euler-Lagrange equation (2.3.5), we have

$$-\mu\Delta u^{k'} + K^\top \mathrm{diag}(\lambda)(Ku^{k'} - z) + \alpha\nabla^\top \left(\max(|\nabla u^{k'}|, \gamma^{k'})^{q-2}\nabla u^{k'}\right) = 0. \tag{2.3.6}$$

Let $k' \to \infty$ so that $\gamma^{k'} \to 0^+$. On the active set $\mathcal{A}(u^*)$ where $|\nabla u^*| > 0$, the first argument of the max-function in (2.3.6) is taken in the limit, i.e.

$$-\mu\Delta u^* + K^\top \mathrm{diag}(\lambda)(Ku^* - z) + \alpha\nabla^\top \left(|\nabla u^*|^{q-2}\nabla u^*\right) = 0, \text{ for } (i,j) \in \mathcal{A}(u^*).$$

On the inactive set $\mathcal{I}(u^*)$, we have $\nabla u^* = 0$ by definition. Thus we conclude that u^* satisfies the Euler-Lagrange equation (2.3.3). □

In particular, if each u^k is a global minimizer of the Huberized problem (2.3.4), with an analogous argument to the proof of Theorem 2.3.1 we have the coercivity of all (f_{γ^k}), uniformly with respect to γ^k. Therefore the sequence (u^k) is uniformly bounded, and the same conclusion as in Theorem 2.3.2 can be drawn.

Corollary 2.3.3. *Let the assumptions in Theorem 2.3.1 hold true. Further assume that* (u^k) *is a sequence such that each* u^k *is a global minimizer of the Huberized problem (2.3.4). Then as* $\gamma^k \to 0^+$*, there exists a subsequence of* (u^k) *converging to some* $u^* \in \mathbb{R}^{|\Omega|}$*, which satisfies the original Euler-Lagrange equation (2.3.3).*

We note that finding global minimizers for nonconvex problems often represents a challenging (if not impossible) task. Therefore, our next task is to design and analyze an algorithm for numerically finding local minimizers of (2.3.4).

We start by noting that the gradient mapping in (2.3.5), i.e. $\nabla f_\gamma : \mathbb{R}^{|\Omega|} \to \mathbb{R}^{|\Omega|}$, is locally Lipschitz. According to Rademacher's Theorem, ∇f_γ is differentiable almost everywhere. Then the generalized Hessian of f_γ at u [Cla83], denoted by $\partial^2 f_\gamma(u)$, is defined as the convex hull of $\partial_B^2 f_\gamma(u)$, where $\partial_B^2 f_\gamma(u)$ consists of all matrices in $\mathbb{R}^{|\Omega|\times|\Omega|}$ that are limits of sequences of the form $\nabla^2 f_\gamma(u^k)$ with $u^k \to u$ and ∇f_γ differentiable at all u^k, i.e.

$$\partial_B^2 f_\gamma(u) := \{\lim \nabla^2 f_\gamma(u^k) : u^k \to u,\ \nabla f_\gamma \text{ is differentiable at } u^k\}.$$

Moreover, the gradient mapping $\nabla f_\gamma : \mathbb{R}^{|\Omega|} \to \mathbb{R}^{|\Omega|}$ is semismooth at any u, i.e.

$$\lim_{\substack{V(u+td') \in \partial^2 f_\gamma(u+td'),\\ d' \to d,\ t \to 0^+}} V(u+td')d' \text{ exists for all } d \in \mathbb{R}^{|\Omega|};$$

see [QS93]. Due to Theorem 2.3 in [QS93], ∇f_γ is directionally differentiable at any u, and for any $V(u+d) \in \partial^2 f_\gamma(u+d)$,

$$\|V(u+d)d - \nabla^2 f_\gamma(u;d)\| = o(\|d\|), \text{ as } \|d\| \to 0,$$

where $o(t)/t \to 0$ as $t \to 0^+$, and $\nabla^2 f_\gamma(u;d)$ denotes the directional derivative of ∇f_γ at u in direction d. Thus, for any $V(u+d) \in \partial_B^2 f_\gamma(u+d)$ we have

$$\|\nabla f_\gamma(u+d) - \nabla f_\gamma(u) - V(u+d)d\| = o(\|d\|), \text{ as } \|d\| \to 0. \tag{2.3.7}$$

In our subsequently defined algorithm, we are in particular interested in the elements of the (possibly) set-valued mapping $\partial_B^2 f_\gamma$ at u, which can be written explicitly as follows:

$$\begin{aligned}\nabla_B^2 f_\gamma(u) := &-\mu\Delta + K^\top \mathrm{diag}(\lambda)K\\ &+ \alpha\nabla^\top \mathrm{diag}\left(\max(|\nabla u|,\gamma)^{q-2}(I - (2-q)\chi_{\mathcal{A}}(u)\max(|\nabla u|,\gamma)^{-2}(\nabla u)(\nabla u)^\top)\right)\nabla,\end{aligned}$$

where $\chi_{\mathcal{A}}(u)$ is defined by

$$(\chi_{\mathcal{A}}(u))_{ij} := \begin{cases} 1, & \text{if } |(\nabla u)_{ij}| \geq \gamma,\\ 0, & \text{otherwise.}\end{cases}$$

We shall refer to $\nabla_B^2 f_\gamma(u)$ as the *B-Hessian* of f at u.

Due to its favorable local convergence properties, we are interested in applying a generalized version of Newton's method for solving (2.3.5). In variational image processing it has turned out that primal-dual Newton schemes are typically superior to purely primal or dual iterations; see, e.g., [CGM99, HK04, HS06]. Hence, we reformulate the Euler-Lagrange equation (2.3.5) by introducing a new variable $p \in (\mathbb{R}^{|\Omega|})^2$, which plays the role of a dual variable, i.e.

$$\begin{cases} -\mu\Delta u + K^\top \mathrm{diag}(\lambda)(Ku - z) + \alpha\nabla^\top p = 0,\\ \max(|\nabla u|,\gamma)^{2-q} p = \nabla u.\end{cases} \tag{2.3.8}$$

This system is the starting point for developing our semismooth Newton scheme in the next section.

2.4 Primal-dual Newton method

2.4.1 Regularized Newton via reweighted Euler-Lagrange equation

In order to handle the nonlinear diffusion term (which contains the $(q-2)$-th power of the max-term) in the Euler-Lagrange equation (2.3.5), we invoke an approach relying on reweighting. Similar techniques were previously considered in [VO96, CM99, NC07, CY08, DDFG10]. In fact, let u^k be our current approximation of a solution to (2.3.5). Then the reweighted Euler-Lagrange equation is given by

$$-\mu\Delta u + K^\top \mathrm{diag}(\lambda)(Ku-z) + \alpha\nabla^\top \left(w^k \max(|\nabla u|,\gamma)^{-r}\nabla u\right) = 0, \tag{2.4.1}$$

with $0 \le r \le 2-q$ and the weight w^k defined by

$$w^k := \max(|\nabla u^k|,\gamma)^{q+r-2}.$$

We further introduce a reweighted dual variable

$$p = w^k \max(|\nabla u|,\gamma)^{-r}\nabla u.$$

As a result, the equation (2.3.8) may be written as

$$\begin{cases} -\mu\Delta u + K^\top \mathrm{diag}(\lambda)(Ku-z) + \alpha\nabla^\top p = 0, \\ (w^k)^{-1}\max(|\nabla u|,\gamma)^r p = \nabla u. \end{cases} \tag{2.4.2}$$

Next, at u^k we define the active set $\mathcal{A}^k := \{(i,j)\in\Omega : |(\nabla u^k)_{ij}| \ge \gamma\}$. Given a current approximation (u^k,p^k), we apply a generalized linearization to (2.4.2) and obtain the semismooth Newton system

$$\begin{bmatrix} -\mu\Delta + K^\top\mathrm{diag}(\lambda)K & \alpha\nabla^\top \\ -\widetilde{C}^k(r)\nabla & \mathrm{diag}((m^k)^{2-q}e) \end{bmatrix} \begin{bmatrix} \delta u^{k+1} \\ \delta p^{k+1} \end{bmatrix} = \begin{bmatrix} \mu\Delta u^k - K^\top\mathrm{diag}(\lambda)(Ku^k - z) - \alpha\nabla^\top p^k \\ \nabla u^k - (m^k)^{2-q}p^k \end{bmatrix}, \tag{2.4.3}$$

where

$$m^k := \max(|\nabla u^k|,\gamma), \tag{2.4.4}$$

$$\widetilde{C}^k(r) := I - r\mathrm{diag}(\chi_{\mathcal{A}^k}(m^k)^{-q}p^k)\begin{bmatrix} \mathrm{diag}(\nabla_x u^k) & \mathrm{diag}(\nabla_y u^k) \\ \mathrm{diag}(\nabla_x u^k) & \mathrm{diag}(\nabla_y u^k) \end{bmatrix}. \tag{2.4.5}$$

After eliminating δp^{k+1}, we are left with the linear system

$$\widetilde{H}^k(r)\delta u^{k+1} = -g^k, \tag{2.4.6}$$

where

$$\widetilde{H}^k(r) := -\mu\Delta + K^\top\mathrm{diag}(\lambda)K + \alpha\nabla^\top\mathrm{diag}((m^k)^{q-2}e)\widetilde{C}^k(r)\nabla, \tag{2.4.7}$$

$$g^k := -\mu\Delta u^k + K^\top\mathrm{diag}(\lambda)(Ku^k - z) + \alpha\nabla^\top((m^k)^{q-2}\nabla u^k). \tag{2.4.8}$$

Note that $g^k = \nabla f_\gamma(u^k)$ in (2.3.5). Upon solving (2.4.6) for δu^{k+1}, we compute δp^{k+1} according to (2.4.3), i.e.

$$\delta p^{k+1} = (m^k)^{q-2}(\nabla u^k + \widetilde{C}^k(r)\nabla \delta u^{k+1}) - p^k. \tag{2.4.9}$$

Assuming that δu^{k+1} is a descent direction for f_γ at u^k, i.e. $(g^k)^\top \delta u^{k+1} < 0$, we update $u^{k+1} := u^k + a^k \delta u^{k+1}$ and $p^{k+1} := p^k + a^k \delta p^{k+1}$ with a suitable step size a^k, and then go to the next Newton iteration.

Note that $H^k := \widetilde{H}^k(2-q)$ is the B-Hessian in the non-reweighted primal-dual Newton method [VO96, HS06]. We observe that the reweighting procedure is, in fact, equivalent to a regularization of the B-Hessian of the non-reweighting approach, which we call the *R-regularization* in our presentation. In order to see this, let

$$R^k := \alpha \nabla^\top \mathrm{diag}(\chi_{\mathcal{A}^k}(m^k)^{-2}p^k) \begin{bmatrix} \mathrm{diag}(\nabla_x u^k) & \mathrm{diag}(\nabla_y u^k) \\ \mathrm{diag}(\nabla_x u^k) & \mathrm{diag}(\nabla_y u^k) \end{bmatrix} \nabla.$$

Then the Newton system (2.4.6) becomes

$$(H^k + \beta R^k)\delta u^{k+1} = -g^k, \tag{2.4.10}$$

with $\beta = 2 - q - r$.

Subsequently we consider variable β, i.e. $\beta = \beta^k$, and a slight modification of the R-matrix to guarantee (i) well-definedness of the Newton iteration defined below, (ii) the aforementioned descent property and (iii) ultimately the local superlinear convergence of our overall algorithmic scheme. For the latter, we show in the proof of Theorem 2.4.10 that $\lim_{k\to\infty} \beta^k = 0$. Thus, the R-regularization vanishes for $k \to \infty$.

2.4.2 Infeasible Newton technique

We next study feasibility properties of the iterates of a semismooth Newton method relying on (2.4.6) and definiteness of $\widetilde{H}^k(r)$. For this discussion, we return to the reweighted Euler-Lagrange equation (2.4.1) with $0 \le r \le 1$ (or $1 - q \le \beta \le 2 - q$). In particular, assuming that $p^k = |\nabla u^k|^{q-2}\nabla u^k$ on $\mathcal{A}^k$, we have that

$$\widetilde{C}^k(r) = I - r\mathrm{diag}(\chi_{\mathcal{A}^k}(m^k)^{-2}e) \begin{bmatrix} \mathrm{diag}(|\nabla_x u^k|^2) & \mathrm{diag}(\nabla_x u^k \nabla_y u^k) \\ \mathrm{diag}(\nabla_x u^k \nabla_y u^k) & \mathrm{diag}(|\nabla_y u^k|^2) \end{bmatrix} \succeq 0, \tag{2.4.11}$$

where "$\succeq$" indicates positive semidefiniteness of a matrix. Therefore, we conclude

$$\widetilde{H}^k(r) = -\mu\Delta + K^\top \mathrm{diag}(\lambda)K + \alpha\nabla^\top \mathrm{diag}((m^k)^{q-2}e)\widetilde{C}^k(r)\nabla \succ 0,$$

i.e. $\widetilde{H}^k(r)$ is positive definite, since $-\mu\Delta + K^\top \mathrm{diag}(\lambda)K \succ 0$ under the hypothesis (2.3.2). In general, however, $\widetilde{H}^k(r)$ may be indefinite during semismooth Newton iterations.

In the following, we derive a sufficient condition for r (or β) such that the system matrix $\widetilde{H}^k(r)$ is positive definite; see Theorem 2.4.2 below. This property of $\widetilde{H}^k(r)$ is useful to guarantee that a descent direction δu^k is computed in each Newton iteration. Moreover, it constitutes an iteration dependent regularization scheme.

For this purpose, we propose two modifications of the system matrix $\widetilde{H}^k(r)$. First, we replace p^k by p_+^k, where

$$p_+^k := \frac{\chi_{\mathcal{A}^k}(m^k)^{q-1}p^k}{\max((m^k)^{q-1}, |p^k|)} + (1-\chi_{\mathcal{A}^k})p^k.$$

Note that the modified p_+^k satisfies its feasibility condition on $\mathcal{A}^k$, i.e.

$$|(p_+^k)_{ij}| \le |(\nabla u^k)_{ij}|^{q-1}, \quad \text{whenever } (i,j) \in \mathcal{A}^k. \tag{2.4.12}$$

Secondly, we replace $\widetilde{C}^k(r)$ by its symmetrization denoted by $\widetilde{C}_+^k(r)$, i.e.

$$\begin{aligned}\widetilde{C}_+^k(r) :=& \frac{\widetilde{C}^k(r) + \widetilde{C}^k(r)^\top}{2} = I - r\mathrm{diag}(\chi_{\mathcal{A}^k}(m^k)^{-q})\cdot \\ &\cdot \begin{bmatrix} \mathrm{diag}((p_+^k)_x \nabla_x u^k) & \mathrm{diag}(\frac{1}{2}((p_+^k)_x\nabla_y u^k + (p_+^k)_y \nabla_x u^k)) \\ \mathrm{diag}(\frac{1}{2}((p_+^k)_x\nabla_y u^k + (p_+^k)_y \nabla_x u^k)) & \mathrm{diag}((p_+^k)_y \nabla_y u^k)\end{bmatrix}.\end{aligned} \tag{2.4.13}$$

Accordingly, the system matrix $\widetilde{H}^k(r)$ in (2.4.6) is replaced by $\widetilde{H}_+^k(r)$ with

$$\widetilde{H}_+^k(r) := -\mu\Delta + K^\top \mathrm{diag}(\lambda)K + \alpha\nabla^\top \mathrm{diag}((m^k)^{q-2}e)\widetilde{C}_+^k(r)\nabla, \tag{2.4.14}$$

and the regularizer R^k is replaced by R_+^k with

$$\begin{aligned}R_+^k :=& \alpha\nabla^\top \mathrm{diag}(\chi_{\mathcal{A}^k}(m^k)^{-2})\cdot \\ &\cdot \begin{bmatrix} \mathrm{diag}((p_+^k)_x \nabla_x u^k) & \mathrm{diag}(\frac{1}{2}((p_+^k)_x\nabla_y u^k + (p_+^k)_y \nabla_x u^k)) \\ \mathrm{diag}(\frac{1}{2}((p_+^k)_x\nabla_y u^k + (p_+^k)_y \nabla_x u^k)) & \mathrm{diag}((p_+^k)_y \nabla_y u^k)\end{bmatrix}\nabla.\end{aligned} \tag{2.4.15}$$

Lemma 2.4.1. *Assume that* $0 \le r \le 1$ *(or equivalently* $1-q \le \beta \le 2-q$*) and the feasibility condition (2.4.12) holds true. Then the matrix* $\widetilde{C}_+^k(r)$ *given in (2.4.13) is positive semidefinite.*

Proof. By reordering, it suffices to show that each 2-by-2 block

$$[\widetilde{C}_+^k(r)]_{ij} = I - r\chi_{\mathcal{A}^k}(m^k)^{-q}\begin{bmatrix} (p_+^k)_x \nabla_x u^k & \frac{1}{2}((p_+^k)_x\nabla_y u^k + (p_+^k)_y \nabla_x u^k) \\ \frac{1}{2}((p_+^k)_x\nabla_y u^k + (p_+^k)_y \nabla_x u^k) & (p_+^k)_y \nabla_y u^k\end{bmatrix}$$

is positive semidefinite. For the ease of notation, the subscripts ij are frequently omitted for the remainder of this proof.

We distinguish two cases with respect to (i,j). First, consider the case where $(i,j) \notin \mathcal{A}^k$. Then we have $[\widetilde{C}_+^k(r)]_{ij} = I$ and the assertion holds immediately.

In the second case where $(i,j) \in \mathcal{A}^k$, we have

$$[\widetilde{C}_+^k(r)]_{ij} = \begin{bmatrix} 1 - r|\nabla u^k|^{-q}(p_+^k)_x \nabla_x u^k & -\frac{r}{2}|\nabla u^k|^{-q}((p_+^k)_x\nabla_y u^k + (p_+^k)_y \nabla_x u^k) \\ -\frac{r}{2}|\nabla u^k|^{-q}((p_+^k)_x\nabla_y u^k + (p_+^k)_y \nabla_x u^k) & 1 - r|\nabla u^k|^{-q}(p_+^k)_y \nabla_y u^k\end{bmatrix}.$$

This 2-by-2 block has nonnegative eigenvalues, since its diagonal elements are nonnegative and its determinant satisfies

$$\begin{aligned}
&(1-r|\nabla u^k|^{-q}(p_+^k)_x\nabla_x u^k)(1-r|\nabla u^k|^{-q}(p_+^k)_y\nabla_y u^k)-\frac{r^2}{4}|\nabla u^k|^{-2q}|(p_+^k)_x\nabla_y u^k+(p_+^k)_y\nabla_x u^k|^2\\
=&1-r|\nabla u^k|^{-q}((p_+^k)_x\nabla_x u^k+(p_+^k)_y\nabla_y u^k)-\frac{r^2}{4}|\nabla u^k|^{-2q}|(p_+^k)_x\nabla_y u^k-(p_+^k)_y\nabla_x u^k|^2\\
=&1-r|\nabla u^k|^{-q}((p_+^k)_x\nabla_x u^k+(p_+^k)_y\nabla_y u^k)-\frac{r^2}{4}|\nabla u^k|^{-2q}\cdot\\
&\cdot\Big[(|(p_+^k)_x|^2+|(p_+^k)_y|^2)(|\nabla_x u^k|^2+|\nabla_y u^k|^2)-|(p_+^k)_x\nabla_x u^k+(p_+^k)_y\nabla_y u^k|^2\Big]\\
=&-\frac{r^2}{4}|\nabla u^k|^{2-2q}|p_+^k|^2+\Big[1-\frac{r}{2}|\nabla u^k|^{-q}((p_+^k)_x\nabla_x u^k+(p_+^k)_y\nabla_y u^k)\Big]^2\\
\geq&-\frac{r^2}{4}|\nabla u^k|^{2-2q}|p_+^k|^2+\Big[1-\frac{r}{2}|\nabla u^k|^{1-q}|p_+^k|\Big]^2=1-r|\nabla u^k|^{1-q}|p_+^k|\geq 0.
\end{aligned}$$

In deriving the above inequalities, we have used the assumption that $0 \leq r \leq 1$, the feasibility condition (2.4.12), and the Cauchy-Schwarz inequality. □

The following theorem is an immediate consequence of Lemma 2.4.1 and the structure of $\widetilde{H}_+^k(r)$.

Theorem 2.4.2 (Sufficient condition for descent property)**.** *Suppose the assumptions of Lemma 2.4.1 are satisfied. Then the following statements hold true:*

1. *The matrix $\widetilde{H}_+^k(r)$ is positive definite.*
2. *We have the following estimate on the spectrum of $\widetilde{H}_+^k(r)$:*

$$\begin{aligned}
&\lambda_{\min}(\widetilde{H}_+^k(r))\geq\lambda_{\min}(-\mu\Delta+K^\top\mathrm{diag}(\lambda)K),\\
&\lambda_{\max}(\widetilde{H}_+^k(r))\leq\lambda_{\max}(-(\mu+3\alpha\gamma^{q-2})\Delta+K^\top\mathrm{diag}(\lambda)K).
\end{aligned}$$

3. *We obtain from (2.4.6) a descent direction δu^{k+1} satisfying*

$$-\frac{(g^k)^\top\delta u^{k+1}}{\|g^k\|\|\delta u^{k+1}\|}\geq\frac{\lambda_{\min}(\widetilde{H}_+^k(r))}{\lambda_{\max}(\widetilde{H}_+^k(r))}\geq\bar{\epsilon}_d:=\frac{\lambda_{\min}(-\mu\Delta+K^\top\mathrm{diag}(\lambda)K)}{\lambda_{\max}(-(\mu+3\alpha\gamma^{q-2})\Delta+K^\top\mathrm{diag}(\lambda)K)}.$$

2.4.3 Superlinear convergence by adaptive regularization

Using the results in [VO96, CM99], one readily finds that the R-regularized version of the semismooth Newton method with fixed β, which results in the reweighting approach, is linearly convergent.

In this subsection, we propose a new adaptively R-regularized version of the semismooth Newton method that attains superlinear local convergence. This requires an appropriate update strategy for $\beta > 0$. For this purpose, we propose a trust-region-type scheme; see, e.g., [NW06, CGT00] for comprehensive discussions of trust-region methods. Given a current iterate u^k, these methods typically model f_γ locally by a quadratic function $h^k : \mathbb{R}^{|\Omega|} \to \mathbb{R}$ with

$$h^k(d) := f_\gamma(u^k) + (g^k)^\top d + \frac{1}{2} d^\top H_+^k d. \tag{2.4.16}$$

Here we let $H_+^k := \widetilde{H}_+^k(2-q)$; see (2.4.14). Consider now the minimization of h^k subject to the trust-region constraint, i.e.

$$\min \quad h^k(d) \quad \text{over } d \in \mathbb{R}^{|\Omega|} \tag{2.4.17}$$

$$\text{s.t.} \quad \frac{1}{2} d^\top R_{+,\varepsilon}^k d \leq \frac{1}{2}(\sigma^k)^2. \tag{2.4.18}$$

Here $\sigma^k > 0$ represents the trust-region radius, and

$$R_{+,\varepsilon}^k := R_+^k + \varepsilon I, \tag{2.4.19}$$

is defined with an arbitrarily fixed regularization parameter $0 < \varepsilon \ll \alpha$. The existence of a solution to (2.4.17)–(2.4.18) hinges on the interplay of H_+^k and $R_{+,\varepsilon}^k$.

Lemma 2.4.3. *The matrix H_+^k is positive definite on $\{d \in \mathbb{R}^{|\Omega|} : d^\top R_{+,\varepsilon}^k d \leq 0\}$.*

Proof. Suppose $d \in \mathbb{R}^{|\Omega|}$ satisfies $d \neq 0$ and $d^\top R_+^k d \leq -\varepsilon \|d\|^2 < 0$. Then we have

$$d^\top H_+^k d = d^\top(-\mu\Delta + K^\top \mathrm{diag}(\lambda) K)d + \alpha(\nabla d)^\top \mathrm{diag}((m^k)^{q-2} e)\nabla d - (2-q) d^\top R_+^k d > 0,$$

which proves the assertion. □

Theorem 2.4.4. *There exists a solution to the trust-region subproblem (2.4.17)–(2.4.18).*

Proof. Note that the objective is at most quadratic and the feasible set is nonempty and closed. It suffices to show that $h^k(d^l) \to \infty$ for any feasible sequence (d^l) with $\|d^l\| \to \infty$. We shall prove this by contradiction. Let such a sequence (d^l) be given, and assume oppositely that $(h^k(d^l))$ is uniformly bounded from above. For each l, we write $d^l = s^l v^l$ such that $s^l \geq 0$, $v^l \in \mathbb{R}^{|\Omega|}$, and $\|v^l\| = 1$. By compactness, there exists a subsequence of (v^l), say $(v^{l'})$, such that $v^{l'} \to v^*$ for some $v^* \in \mathbb{R}^{|\Omega|}$. The constraint (2.4.18) yields that $(v^{l'})^\top R_{+,\varepsilon}^k v^{l'} \leq (\sigma^k)^2/(s^{l'})^2$. Letting $l' \to \infty$, we get $(v^*)^\top R_{+,\varepsilon}^k v^* \leq 0$. It follows from Lemma 2.4.3 that $(v^*)^\top H_+^k v^* > 0$. Thus we must have $h^k(d^{l'}) \to \infty$ as $l' \to \infty$, which contradicts our assumption. □

Given the current iterate u^k, we aim to determine a search direction d^k by approximately solving the trust-region subproblem. A classical argument in the convergence analysis of trust-region methods requires that the search direction d^k yields a reduction in the model function h^k proportional to the decrease implied by the *Cauchy point* [CGT00].

The Cauchy point is defined by $d_C^k := -t^k g^k$, where t^k minimizes the one-dimensional problem

$$t^k := \arg\min\{h^k(-tg^k) : \ t^2 (g^k)^\top R_{+,\varepsilon}^k g^k \leq (\sigma^k)^2, \ t \geq 0\}.$$

Let $t_*^k := \|g^k\|^2/((g^k)^\top H_+^k g^k)$ be the critical point provided that it exists. The Cauchy point can be explicitly computed through the following three cases:

1. Suppose $(g^k)^\top H_+^k g^k \le 0$. By Lemma 2.4.3, we have $(g^k)^\top R_{+,\varepsilon}^k g^k > 0$. The Cauchy point lies on the boundary of the trust region, i.e. $d_C^k = -\left(\sigma^k / \sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}\right) g^k$, and
$$h^k(0) - h^k(d^k) = \frac{\sigma^k \|g^k\|^2}{\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}} - \frac{(\sigma^k)^2 (g^k)^\top H_+^k g^k}{2(g^k)^\top R_{+,\varepsilon}^k g^k} \ge \frac{\sigma^k \|g^k\|^2}{\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}}. \tag{2.4.20}$$

2. Suppose $(g^k)^\top H_+^k g^k > 0$ and $(t_*^k)^2 (g^k)^\top R_{+,\varepsilon}^k g^k \le (\sigma^k)^2$. Then we have $d_C^k = -t_*^k g^k = -\left(\|g^k\|^2/((g^k)^\top H_+^k g^k)\right) g^k$, and
$$h^k(0) - h^k(d^k) = \frac{\|g^k\|^4}{2(g^k)^\top H_+^k g^k} \ge \frac{\|g^k\|^2}{2\lambda_{\max}(H_+^k)}. \tag{2.4.21}$$

3. Suppose $(g^k)^\top H_+^k g^k > 0$ and $(t_*^k)^2 (g^k)^\top R_{+,\varepsilon}^k g^k > (\sigma^k)^2$. Then similar to the first case, we have $d_C^k = -\left(\sigma^k / \sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}\right) g^k$. In particular, $\sigma^k((g^k)^\top H_+^k g^k)/\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k} < \|g^k\|^2$. Therefore, we have
$$h^k(0) - h^k(d^k) = \frac{\sigma^k \|g^k\|^2}{\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}} - \frac{(\sigma^k)^2 (g^k)^\top H_+^k g^k}{2(g^k)^\top R_{+,\varepsilon}^k g^k} \ge \frac{\sigma^k \|g^k\|^2}{2\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}}. \tag{2.4.22}$$

The search direction d^k is said to satisfy the *Cauchy-point-based model reduction criterion* if
$$h^k(0) - h^k(d^k) \ge C\|g^k\|^2 \eta^k, \tag{2.4.23}$$
for some constant $C > 0$, where
$$\eta^k := \begin{cases} \dfrac{\sigma^k}{\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}}, & \text{if } (g^k)^\top H_+^k g^k \le 0, \\ \dfrac{1}{\lambda_{\max}(H_+^k)}, & \text{if } (g^k)^\top R_{+,\varepsilon}^k g^k \le 0, \\ \min\left(\dfrac{\sigma^k}{\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}}, \dfrac{1}{\lambda_{\max}(H_+^k)}\right), & \text{otherwise.} \end{cases} \tag{2.4.24}$$

Due to Lemma 2.4.3, η^k is well-defined. It is easily seen that (2.4.20)–(2.4.22) satisfy the criterion (2.4.23) with $C = 1/2$.

Now we turn to the computation of an approximate solution to the trust-region subproblem (2.4.17)–(2.4.18). In the forthcoming Theorem 2.4.5, we shall characterize this solution d_*^k by
$$(H_+^k + \beta_*^k R_{+,\varepsilon}^k) d_*^k = -g^k, \tag{2.4.25}$$
$$\beta_*^k \left((d_*^k)^\top R_{+,\varepsilon}^k d_*^k - (\sigma^k)^2\right) = 0, \tag{2.4.26}$$
$$H_+^k + \beta_*^k R_{+,\varepsilon}^k \succeq 0, \tag{2.4.27}$$
for some $\beta_*^k \ge 0$. Its proof essentially adopts the proof of Theorem 4.1 in [NW06] under our context.

Theorem 2.4.5. *The trust-region subproblem (2.4.17)–(2.4.18) has a global solution d^k_* if and only if d^k_* is feasible and there exists a scalar $\beta^k_* \geq 0$ such that (2.4.25)–(2.4.27) are satisfied.*

Proof. (if part) Suppose there exists $\beta^k_* \geq 0$ such that (2.4.25)–(2.4.27) hold. Then by Lemma 4.7 in [NW06], d^k_* minimizes $\widehat{h}^k : \mathbb{R}^{|\Omega|} \to \mathbb{R}$, where

$$\widehat{h}^k(d^k) := (g^k)^\top d^k + \frac{1}{2}(d^k)^\top (H^k_+ + \beta^k_* R^k_{+,\varepsilon})d^k = h^k(d^k) + \frac{\beta^k_*}{2}(d^k)^\top R^k_{+,\varepsilon} d^k - f_\gamma(u^k).$$

If follows from $\widehat{h}^k(d^k) \geq \widehat{h}^k(d^k_*)$ that

$$\begin{aligned} h^k(d^k) \geq {} & h^k(d^k_*) + \frac{\beta^k_*}{2}((d^k_*)^\top R^k_{+,\varepsilon} d^k_* - (d^k)^\top R^k_{+,\varepsilon} d^k) \\ = & h^k(d^k_*) + \frac{\beta^k_*}{2}((\sigma^k)^2 - (d^k)^\top R^k_{+,\varepsilon} d^k) \geq h^k(d^k_*). \end{aligned}$$

Since d^k is arbitrary but feasible, the assertion follows.

(only-if part) Suppose now that d^k_* is the global solution of the trust-region subproblem (2.4.17)–(2.4.18).

- Case 1: $(d^k_*)^\top R^k_{+,\varepsilon} d^k_* < (\sigma^k)^2$. The second-order necessary conditions of the unconstrained problem imply that

  $$\begin{aligned} \nabla h^k(d^k_*) &= H^k_+ d^k_* + g^k = 0, \\ \nabla^2 h^k(d^k_*) &= H^k_+ \succeq 0. \end{aligned}$$

  We get the desired conclusion with $\beta^k_* = 0$.

- Case 2: $(d^k_*)^\top R^k_{+,\varepsilon} d^k_* = (\sigma^k)^2$. In particular we have $R^k_{+,\varepsilon} d^k_* \neq 0$, and therefore the linear independence constraint qualification (see, e.g., [NW06]) is fulfilled at d^k_*. By the second-order necessary condition, there exists $\beta^k_* \geq 0$ such that

  $$H^k_+ d^k_* + g^k + \beta^k_* R^k_{+,\varepsilon} d^k_* = 0, \tag{2.4.28}$$

  and

  $$v^\top (H^k_+ + \beta^k_* R^k_{+,\varepsilon})v \geq 0, \tag{2.4.29}$$

  for any nonzero vector $v \in \mathbb{R}^{|\Omega|}$ with $v^\top R^k_{+,\varepsilon} d^k_* = 0$.

  It remains to show (2.4.29) for any nonzero vector v with $v^\top R^k_{+,\varepsilon} d^k_* \neq 0$. Let such a vector v be given. In particular we have $R^k_{+,\varepsilon} v \neq 0$. Define

  $$d^k := d^k_* - \frac{2v^\top R^k_{+,\varepsilon} d^k_*}{v^\top R^k_{+,\varepsilon} v} v. \tag{2.4.30}$$

  Then it is easy to check that $(d^k)^\top R^k_{+,\varepsilon} d^k = (\sigma^k)^2$. Since $h^k(d^k) \geq h^k(d^k_*)$, we have

  $$h^k(d^k) \geq h^k(d^k_*) + \frac{\beta^k_*}{2}((d^k_*)^\top R^k_{+,\varepsilon} d^k_* - (d^k)^\top R^k_{+,\varepsilon} d^k).$$

From this and (2.4.28), we infer

$$\frac{1}{2}(d^k - d_*^k)^\top (H_+^k + \beta_*^k R_{+,\varepsilon}^k)(d^k - d_*^k) \geq 0.$$

Thus in view of (2.4.30) we have shown (2.4.29) for any nonzero vector v with $v^\top R_{+,\varepsilon}^k d_*^k \neq 0$, which completes the proof.

□

Based on the above observation concerning h^k and using a complementarity function (see, e.g., [HIK03]), we can equivalently formulate (2.4.25)–(2.4.27), with an arbitrarily fixed scalar $c > 0$, as follows:

$$(H_+^k + \beta_*^k R_{+,\varepsilon}^k)d_*^k = -g^k, \tag{2.4.31}$$

$$\beta_*^k - \max\left(\beta_*^k + \frac{1}{2c}((d_*^k)^\top R_{+,\varepsilon}^k d_*^k - (\sigma^k)^2), 0\right) = 0, \tag{2.4.32}$$

$$H_+^k + \beta_*^k R_{+,\varepsilon}^k \succeq 0. \tag{2.4.33}$$

From this formulation, we propose an adaptively regularized Newton iteration which converges globally and locally at a superlinear rate.

Algorithm 2.4.6 (Adaptively regularized Newton method)**.**

```
Require: input parameters $1 - q \leq \beta_{\max} \leq 2 - q$, $c > 0$, $0 < \rho_1 \leq \rho_2 < 1$, $0 < \kappa_1 < 1 < \kappa_2$,
    $0 < \varepsilon \ll \alpha$, $0 < \epsilon_d \leq \bar{\epsilon}_d$, $0 < \tau_1 < 1/2$, $\tau_1 < \tau_2 < 1$.
 1: Initialize the primal and dual variables $(u^0, p^0)$, the regularization scalar $\beta^0 \geq 0$, and the
    trust-region radius $\sigma^0 > 0$. Set $k := 0$.
 2: repeat {outer loop}
 3:   Generate $H_+^k$, $R_{+,\varepsilon}^k$, and $g^k$.
 4:   repeat {inner loop}
 5:     Solve $(H_+^k + \beta^k R_{+,\varepsilon}^k)d^k = -g^k$ for $d^k$.
 6:     if $-(g^k)^\top d^k/(\|g^k\|\|d^k\|) < \epsilon_d$ then
 7:       Set $\beta^k := \beta_{\max}$ and return to step 5.
 8:     end if
 9:     if $\beta^k = \beta_{\max}$ and $(d^k)^\top R_{+,\varepsilon}^k d^k > (\sigma^k)^2$ then
10:       Set $\sigma^k := \sqrt{(d^k)^\top R_{+,\varepsilon}^k d^k}$ and go to step 15.
11:     end if
12:     Update $\beta^k := \beta^k + ((d^k)^\top R_{+,\varepsilon}^k d^k - (\sigma^k)^2)/(2c)$.
13:     Project $\beta^k$ onto the interval $[0, \beta_{\max}]$, i.e. set $\beta^k := \max(\min(\beta^k, \beta_{\max}), 0)$.
14:   until the stopping criterion for the inner loop is fulfilled.
15:   Evaluate $\rho^k := [f_\gamma(u^k) - f_\gamma(u^k + d^k)]/[f_\gamma(u^k) - (f_\gamma(u^k) + (g^k)^\top d^k + (d^k)^\top H_+^k d^k/2)]$.
16:   if $\rho^k < \rho_1$ then
```

```
17:        Set $\sigma^{k+1} := \kappa_1 \sigma^k$.
18:     else if $\rho^k > \rho_2$ then
19:        Set $\sigma^{k+1} := \kappa_2 \sigma^k$.
20:     else
21:        $\sigma^{k+1} := \sigma^k$.
22:     end if
23:     Determine the step size $a^k$ along the search direction $d^k$ such that $u^{k+1} = u^k + a^k d^k$
        satisfies the following Wolfe-Powell conditions:
            $f_\gamma(u^{k+1}) \le f_\gamma(u^k) + \tau_1 a^k \nabla f_\gamma(u^k)^\top d^k,$        (2.4.34)
            $\nabla f_\gamma(u^{k+1})^\top d^k \ge \tau_2 \nabla f_\gamma(u^k)^\top d^k.$        (2.4.35)
24:     Set $\delta u^{k+1} := d^k$ and compute $\delta p^{k+1}$ according to (2.4.9). Update $u^{k+1} := u^k + a^k \delta u^{k+1}$
        and $p^{k+1} := p^k + a^k \delta p^{k+1}$.
25:     Initialize the regularization weight $\beta^{k+1} := \beta^k$ for the next iteration.
26:     Set $k := k + 1$.
27:  until the stopping criterion for the outer loop is fulfilled.
```

Concerning the input parameters involved in the above algorithm, we note that these quantities are presented merely for the generality of the algorithm and do not require particular tuning for various imaging restoration tasks. Throughout our numerical experiments in section 2.5, we shall always fix the parameters as follows: $\beta_{\max} = 1.2 - q,\ c = 1,\ \rho_1 = 0.25,\ \rho_2 = 0.75,\ \kappa_1 = 0.25,\ \kappa_2 = 2,\ \varepsilon = 10^{-4}\alpha,\ \epsilon_d = 10^{-8},\ \tau_1 = 0.1,\ \tau_2 = 0.9.$

We observe that Algorithm 2.4.6 combines a trust-region technique for adjusting the weight β in the R-regularization (steps 4–14) with a line search method for updating the iterate along the direction obtained from the approximately weighted R-regularized problem (step 23). We emphasize, however, that the classical trust-region approach might be used instead of the line search procedure for globalizing Newton's method; see section 2.2.3 and the references therein. In Algorithm 2.4.6, the global convergence is guaranteed by the Wolfe-Powell line search while the trust-region-type framework is utilized to guarantee that d^k is a descent direction for f_γ at u^k and to retain the local superlinear convergence of Newton's method. Based on our numerical experience we prefer the Wolfe-Powell line search over other, possibly simpler, rules as it appears to better resolve the line search problem for our nonconvex objective.

Note that our objective f_γ is bounded from below and continuously differentiable. Moreover, its gradient $\nabla f_\gamma(\cdot)$ is locally Lipschitz. Thus Zoutendijk's theorem, recall Theorem 2.2.14, can be applied to derive the global convergence of Algorithm 2.4.6.

Theorem 2.4.7 (Global convergence)**.** *Let $u^{k+1} = u^k + a^k d^k$ such that the Wolfe-Powell conditions (2.8.15)–(2.8.16) are satisfied. Then we have* $\lim_{k\to\infty} \|\nabla f_\gamma(u^k)\| = 0$.

Proof. By Theorem 2.2.14, we have $\sum_{k=0}^{+\infty} \cos^2\theta^k \|g^k\|^2 < +\infty$, where $\cos\theta^k := -\frac{(g^k)^\top d^k}{\|g^k\|\|d^k\|}$. Since

$\cos\theta^k \geq \epsilon_d$ holds true for each k due to steps 6–8 of Algorithm 2.4.6 and Theorem 2.4.2, we conclude that $\lim_{k\to\infty} \|\nabla f_\gamma(u^k)\| = 0$. □

Next we study the local convergence of Algorithm 2.4.6. As a preparatory result, Lemma 2.4.8 investigates the approximation properties of (H_+^k) with respect to $(\nabla_B^2 f_\gamma(u^k))$ and the definiteness properties of $(R_{+,\varepsilon}^k)$.

Lemma 2.4.8. *Assume that the primal-dual sequence (u^k, p^k) converges to some (u^*, p^*) satisfying the Euler-Lagrange system (2.3.8). Then the following statements hold true:*

1. *The modified system matrix H_+^k approaches asymptotically the B-Hessian $\nabla_B^2 f_\gamma(u^k)$, i.e. $\lim_{k\to\infty} \|H_+^k - \nabla_B^2 f_\gamma(u^k)\| = 0$.*

2. *For all sufficiently large k, the matrix $R_{+,\varepsilon}^k$ is strictly positive definite and its minimal eigenvalue satisfies $\lambda_{\min}(R_{+,\varepsilon}^k) > \varepsilon/2$.*

Proof. (Proof of 1.) Let $C^k := \widetilde{C}^k(2-q)$ in (2.4.11) and $C_+^k := \widetilde{C}_+^k(2-q)$ in (2.4.13). For $k \to \infty$ we have $(u^k, p^k) \to (u^*, p^*)$ with the latter satisfying the Euler-Lagrange equation (2.3.8). Further, for all $(i,j) \in \Omega$ we have

$$\begin{aligned} |p_+^k - p^k| &\leq |p^k| \left| \frac{(m^k)^{q-1}}{\max((m^k)^{q-1}, |p^k|)} - 1 \right| \to |p^*| \left| \frac{\max(|\nabla u^*|, \gamma)^{q-1}}{\max(\max(|\nabla u^*|, \gamma)^{q-1}, |p^*|)} - 1 \right| \\ &= |p^*| \left| \frac{\max(|\nabla u^*|, \gamma)^{q-1}}{|p^*| \max(|\nabla u^*|, \gamma)/|\nabla u^*|} - 1 \right| = 0 \end{aligned} \tag{2.4.36}$$

as $k \to \infty$. Moreover, C^k will converge to a symmetric matrix, and therefore $C_+^k = (C^k + (C^k)^\top)/2$ approaches asymptotically C^k, i.e. $\lim_{k\to\infty} \|C_+^k - C^k\| = 0$. Thus, due to the structures of H^k and H_+^k, we have $\lim_{k\to\infty} \|H_+^k - H^k\| = 0$.

Finally, as $(u^k, p^k) \to (u^*, p^*)$, it is easy to see that both H^k and $\nabla_B^2 f_\gamma(u^k)$ converge to $\nabla_B^2 f_\gamma(u^*)$, which yields $\lim_{k\to\infty} \|H^k - \nabla_B^2 f_\gamma(u^k)\| = 0$. Thus we conclude that $\lim_{k\to\infty} \|H_+^k - \nabla_B^2 f_\gamma(u^k)\| = 0$ as desired.

(Proof of 2.) Our proof again utilizes the reordered system as in Lemma 2.4.1. In view of the definition of $R_{+,\varepsilon}^k$, see (2.4.19), and the structure of R_+^k, see (2.4.15), it suffices to show that for all $(i,j) \in \Omega$, the minimal eigenvalue of the 2-by-2 block

$$\chi_{\mathcal{A}^k} (m^k)^{-2} \begin{bmatrix} (p_+^k)_x \nabla_x u^k & \frac{1}{2}((p_+^k)_x \nabla_y u^k + (p_+^k)_y \nabla_x u^k) \\ \frac{1}{2}((p_+^k)_x \nabla_y u^k + (p_+^k)_y \nabla_x u^k) & (p_+^k)_y \nabla_y u^k \end{bmatrix} \tag{2.4.37}$$

goes to zero as $k \to \infty$. The characteristic equation of the 2-by-2 block (2.4.37) without the factor $\chi_{\mathcal{A}^k}$ is given by

$$\begin{aligned} &t^2 - (m^k)^{-2}((p_+^k)_x \nabla_x u^k + (p_+^k)_y \nabla_y u^k) t \\ &\quad + (m^k)^{-4} \left((p_+^k)_x \nabla_x u^k (p_+^k)_y \nabla_y u^k - \frac{1}{4} |(p_+^k)_x \nabla_y u^k + (p_+^k)_y \nabla_x u^k|^2 \right) = 0. \end{aligned}$$

Note that due to (2.4.36) we have $\lim_{k\to\infty} p_+^k = p^*$ such that (u^*, p^*) satisfies (2.3.8). Therefore, as $k \to \infty$, we have

$$\begin{aligned}
&(m^k)^{-4}\left((p_+^k)_x\nabla_x u^k (p_+^k)_y \nabla_y u^k - \frac{1}{4}|(p_+^k)_x\nabla_y u^k + (p_+^k)_y\nabla_x u^k|^2\right)\\
=&-\frac{(m^k)^{-4}}{4}\left[|(p_+^k)_x\nabla_y u^k|^2 + |(p_+^k)_y\nabla_x u^k|^2 - 2(p_+^k)_x(p_+^k)_y\nabla_x u^k\nabla_y u^k\right]\\
=&\frac{(m^k)^{-4}}{4}\left[|p_+^k|^2|\nabla u^k|^2 - |(p_+^k)_x\nabla_x u^k + (p_+^k)_y\nabla_y u^k|^2\right] \to 0, \qquad (2.4.38)
\end{aligned}$$

and

$$(m^k)^{-2}((p_+^k)_x\nabla_x u^k + (p_+^k)_y\nabla_y u^k) \to \max(|\nabla u^*|, \gamma)^{q-4}|\nabla u^*|^2 > 0. \qquad (2.4.39)$$

From (2.4.38) and (2.4.39), we conclude that the minimal eigenvalue of the 2-by-2 block (2.4.37) without the factor $\chi_{\mathcal{A}^k}$ goes to zero as $k \to \infty$. Since $\{\chi_{\mathcal{A}^k}\}$ is uniformly bounded, the minimal eigenvalue of (2.4.37) goes to zero as $(u^k, p^k) \to (u^*, p^*)$, which completes the proof. □

The following lemma verifies the convergence of the inner loop, i.e. steps 4–14 in Algorithm 2.4.6.

Lemma 2.4.9. *Assume that H_+^k and $R_{+,\varepsilon}^k$ are both positive definite, and*

$$0 < \|g^k\| < \sqrt{\frac{c(\lambda_{\min}(H_+^k))^3}{(\lambda_{\max}(R_{+,\varepsilon}^k))^2}}.$$

Then the sequence $\{(\beta_l^k, d_l^k) : l \in \mathbb{N}\}$ generated by the inner iterations, i.e. steps 4–14, of Algorithm 2.4.6 converges to some (β_*^k, d_*^k) satisfying the optimality conditions of the trust-region subproblem; see (2.4.31)–(2.4.33).

Proof. By our assumption, the definiteness condition (2.4.33) is automatically satisfied. In the case where Steps 9–11 of Algorithm 2.4.6 are active, the inner iterations terminate with a modified σ^k such that the conditions (2.4.31)–(2.4.32) are satisfied. Hence, in what follows we assume that Steps 9–11 are inactive all along the sequence $\{(\beta_l^k, d_l^k) : l \in \mathbb{N}\}$.

We define the function $\phi : [0, \beta_{\max}] \to \mathbb{R}$ by

$$\phi(\beta) = \beta + \frac{((H_+^k + \beta R_{+,\varepsilon}^k)^{-1}g^k)^\top R_{+,\varepsilon}^k (H_+^k + \beta R_{+,\varepsilon}^k)^{-1}g^k - (\sigma^k)^2}{2c}.$$

Then by eliminating d^k by $d^k = -(H_+^k + \beta^k R_{+,\varepsilon}^k)^{-1}g^k$ in Step 12 of Algorithm 2.4.6, we have the update rule (Steps 12–13) as follows

$$\beta_{l+1}^k = \max\left(\min\left(\phi(\beta_l^k), \beta_{\max}\right), 0\right).$$

Note that ϕ is continuously differentiable, and its derivative is given by

$$\phi'(\beta) = 1 - \frac{1}{c}(g^k)^\top (H_+^k + \beta R_{+,\varepsilon}^k)^{-1}(R_{+,\varepsilon}^k(H_+^k + \beta R_{+,\varepsilon}^k)^{-1})^2 g^k.$$

It follows from our assumptions that

$$\left|\frac{1}{c}(g^k)^\top (H_+^k + \beta R_{+,\varepsilon}^k)^{-1}(R_{+,\varepsilon}^k (H_+^k + \beta R_{+,\varepsilon}^k)^{-1})^2 g^k\right| \leq \frac{(\lambda_{\max}(R_{+,\varepsilon}^k))^2 \|g^k\|^2}{c\lambda_{\min}(H_+^k + \beta R_{+,\varepsilon}^k)^3}$$
$$\leq \frac{(\lambda_{\max}(R_{+,\varepsilon}^k))^2 \|g^k\|^2}{c(\lambda_{\min}(H_+^k))^3} < 1.$$

By the above inequality and the mean value theorem, there exists a constant $C \in (0,1)$ such that for any $\beta_1, \beta_2 \in [0, \beta_{\max}]$,

$$|\phi(\beta_1) - \phi(\beta_2)| \leq |\beta_1 - \beta_2| \sup_{\beta \in [0,\beta_{\max}]} |\phi'(\beta)| \leq C|\beta_1 - \beta_2|,$$

i.e. ϕ is a *contractive* mapping. As a consequence, the mapping $\beta \mapsto \max(\min(\phi(\beta), \beta_{\max}), 0)$ is also contractive. Thus by the Banach fixed-point theorem, we have $\beta_l^k \to \beta_*^k$ as $l \to \infty$ for some $\beta_*^k \in [0, \beta_{\max}]$. Accordingly, $d_l^k \to d_*^k = -(H_+^k + \beta_*^k R_{+,\varepsilon}^k)^{-1} g^k$ as $l \to \infty$. Moreover, (β_*^k, d_*^k) satisfies (2.4.31)–(2.4.32), which completes the proof. □

Now we are in the position to present our local convergence result.

Theorem 2.4.10 (Local convergence). *Let $\{d^k\}$ be generated by Algorithm 2.4.6, and let the sequence $\{(u^k, p^k)\}$ converge to some (u^*, p^*) satisfying the Euler-Lagrange system (2.3.8). Assume that all elements in $\partial_B^2 f_\gamma(u^*)$ are strictly positive definite. Then Algorithm 2.4.6 is locally superlinearly convergent, i.e. for sufficiently large k we have*

$$\|u^{k+1} - u^*\| = o(\|u^k - u^*\|) \quad \textit{for } k \to \infty. \tag{2.4.40}$$

Proof. Throughout the proof we argue only for sufficiently large k. From our assumption that all elements of $\partial_B^2 f_\gamma(u^*)$ are strictly positive definite, it follows that all elements in $\partial_B^2 f_\gamma(u^k)$, including $\nabla_B^2 f_\gamma(u^k)$, are strictly positive definite with uniformly bounded inverses; see Lemma 2.2.11. Furthermore, due to Lemma 2.4.8 we have that H_+^k is also strictly positive definite.

Since $R_{+,\varepsilon}^k \succ 0$ according to Lemma 2.4.8, we have

$$-(d^k)^\top g^k = (d^k)^\top H_+^k d^k + \beta^k (d^k)^\top R_{+,\varepsilon}^k d^k \geq \lambda_{\min}(H_+^k)\|d^k\|^2 \geq 0.$$

Letting $k \to \infty$, we have $\|d^k\| \to 0$ since $\|g^k\| \to 0$ by Theorem 2.2.14.

Next, we show that $\lim_{k\to\infty} \beta^k = 0$. From the semismoothness property (2.3.7) and Lemma 2.4.8, we have that as $k \to \infty$,

$$|(f_\gamma(u^k) - f_\gamma(u^k + d^k)) - (h^k(0) - h^k(d^k))| = \left| f_\gamma(u^k + d^k) - f_\gamma(u^k) - (g^k)^\top d^k - \frac{1}{2}(d^k)^\top H_+^k d^k \right|$$
$$\leq \left| f_\gamma(u^k + d^k) - f_\gamma(u^k) - (g^k)^\top d^k - \frac{1}{2}(d^k)^\top \nabla_B^2 f_\gamma(u^k) d^k \right| + \left| \frac{1}{2}(d^k)^\top (\nabla_B^2 f_\gamma(u^k) - H_+^k) d^k \right|$$
$$= o(\|d^k\|^2). \tag{2.4.41}$$

For sufficiently large k, all assumptions in Lemma 2.4.9 hold true. Therefore, we have that

$$(d^k)^\top R^k_{+,\varepsilon} d^k \le \nu^2 (\sigma^k)^2,$$

for some constant $\nu > 0$, since otherwise (2.4.32) would fail. Lemma 2.4.9 also implies that d^k will satisfy the Cauchy-point-based model reduction criterion (2.4.23) after sufficiently many inner iterations. In fact, only the last case in (2.4.24) may occur. So as $k \to \infty$, we have

$$\begin{aligned} &h^k(0) - h^k(d^k) \ge C\|g^k\|^2 \min\left(\frac{\sigma^k}{\sqrt{(g^k)^\top R^k_{+,\varepsilon} g^k}}, \frac{1}{\lambda_{\max}(H^k_+)} \right) \\ \ge& C\|g^k\| \min\left(\frac{\|g^k\|\sqrt{(d^k)^\top R^k_{+,\varepsilon} d^k}}{\nu\sqrt{(g^k)^\top R^k_{+,\varepsilon} g^k}}, \frac{\|g^k\|}{\lambda_{\max}(H^k_+)} \right) \ge C\|g^k\| \min\left(\frac{\sqrt{\lambda_{\min}(R^k_{+,\varepsilon})}\|d^k\|}{\nu\sqrt{\lambda_{\max}(R^k_{+,\varepsilon})}}, \frac{\|g^k\|}{\lambda_{\max}(H^k_+)} \right) \\ \ge& C\lambda_{\min}(H^k_+) \min\left(\frac{\sqrt{\lambda_{\min}(R^k_{+,\varepsilon})}}{\nu\sqrt{\lambda_{\max}(R^k_{+,\varepsilon})}}, \frac{\lambda_{\min}(H^k_+)}{\lambda_{\max}(H^k_+)} \right) \|d^k\|^2. \end{aligned} \tag{2.4.42}$$

Combining (2.4.41) and (2.4.42), we have that as $k \to \infty$

$$|\rho^k - 1| = \frac{|(f_\gamma(u^k) - f_\gamma(u^k + d^k)) - (h^k(0) - h^k(d^k))|}{|h^k(0) - h^k(d^k)|} \le o(1) \to 0.$$

Thus the sequence $\{\sigma^k\}$ is uniformly bounded away from 0. Consequently, $\lim_{k\to\infty} \beta^k = 0$, and the Dennis-Moré condition [DM77] is satisfied, i.e. as $k \to \infty$,

$$\frac{\|(H^k_+ + \beta^k R^k_{+,\varepsilon})d^k - \nabla^2_B f_\gamma(u^*) d^k\|}{\|d^k\|} \le \|H^k_+ - \nabla^2_B f_\gamma(u^*)\| + \beta^k \lambda_{\max}(R^k_{+,\varepsilon}) \to 0,$$

as the sequence $\{\lambda_{\max}(R^k_{+,\varepsilon})\}$ is uniformly bounded.

It follows from the semismoothness property (2.3.7) that

$$\begin{aligned} &f_\gamma(u^k + d^k) - f_\gamma(u^k) - \tau_1 \nabla f_\gamma(u^k)^\top d^k = (1-\tau_1)\nabla f_\gamma(u^k)^\top d^k + \frac{1}{2}(d^k)^\top \nabla^2_B f_\gamma(u^k) d^k + o(\|d^k\|^2) \\ =&(d^k)^\top[(\tau_1 - 1)(H^k_+ + \beta^k R^k_{+,\varepsilon}) + \frac{1}{2}\nabla^2_B f_\gamma(u^k)]d^k + o(\|d^k\|^2) \\ =&(\tau_1 - \frac{1}{2})(d^k)^\top \nabla^2_B f_\gamma(u^k) d^k + o(\|d^k\|^2) \le 0, \end{aligned}$$

and

$$\begin{aligned} &\nabla f_\gamma(u^k + d^k)^\top d^k - \tau_2 \nabla f_\gamma(u^k)^\top d^k = (d^k)^\top \nabla^2_B f_\gamma(u^k) d^k + (1-\tau_2)\nabla f_\gamma(u^k)^\top d^k + o(\|d^k\|^2) \\ =&(d^k)^\top[(\tau_2 - 1)(H^k_+ + \beta^k R^k_{+,\varepsilon}) + \nabla^2_B f_\gamma(u^k)]d^k + o(\|d^k\|^2) = \tau_2 (d^k)^\top \nabla^2_B f_\gamma(u^k) d^k + o(\|d^k\|^2) \ge 0, \end{aligned}$$

for sufficiently large k since $\|d^k\| \to 0$ as $k \to \infty$. Hence the Wolfe-Powell conditions (2.8.15)–(2.8.16) are satisfied for $a^k = 1$, i.e. $u^{k+1} = u^k + d^k$, for all sufficiently large k.

Let $d_N^k := -\nabla_B^2 f_\gamma(u^k)^{-1} g^k$. Note that

$$\begin{aligned}
&\|d^k - d_N^k\| = \|\nabla_B^2 f_\gamma(u^k)^{-1}(\nabla_B^2 f_\gamma(u^k) d^k + g^k)\| \\
=&\|\nabla_B^2 f_\gamma(u^k)^{-1}(\nabla_B^2 f_\gamma(u^k) - (H_+^k + \beta^k R_+^k))d^k\| \leq \|\nabla_B^2 f_\gamma(u^k)^{-1}\| o(\|d^k\|) = o(\|d^k\|),
\end{aligned}$$

since $\{\nabla_B^2 f_\gamma(u^k)^{-1}\}$ is uniformly bounded as $u^k \to u^*$ for $k \to \infty$. As a consequence, we have

$$\|u^{k+1} - u^*\| = \|u^k + d^k - u^*\| \leq \|u^k + d_N^k - u^*\| + \|d^k - d_N^k\| = o(\|u^k - u^*\|).$$

We have used that $\|u^k + d_N^k - u^*\| = o(\|u^k - u^*\|)$ (see, e.g., Theorem 8.5 in [IK08]), and that $\|d^k\| = O(\|u^k - u^*\|)$. From this we conclude that Algorithm 2.4.6 is locally superlinearly convergent. □

The assumption of Theorem 2.4.10 relates to second-order sufficient optimality conditions for smooth problems. Although such assumptions typically occur in the optimization literature (also in the context of nonsmooth problems), they are difficult to check in an algorithm.

2.5 Numerical experiments

In this section we present numerical results obtained by our primal-dual Newton method. Throughout this section, Ω denotes the m-by-n pixel-domain, i.e. $\Omega = \{(i,j) \in \mathbb{Z}^2 : 1 \leq i \leq m,\ 1 \leq j \leq n\}$. We discretize the gradient operator by $(\nabla u)_{ij} = \big((u_{i+1,j} - u_{i,j})/\omega, (u_{i,j+1} - u_{i,j})/\omega\big)$ with $\omega = \sqrt{1/|\Omega|}$. We set $u_{ij} = 0$ whenever $(i,j) \notin \Omega$. Unless otherwise specified, the following parameters are used in our experiments: $q = 0.75,\ \mu = 10^{-4}\alpha,\ \gamma = 0.1$.

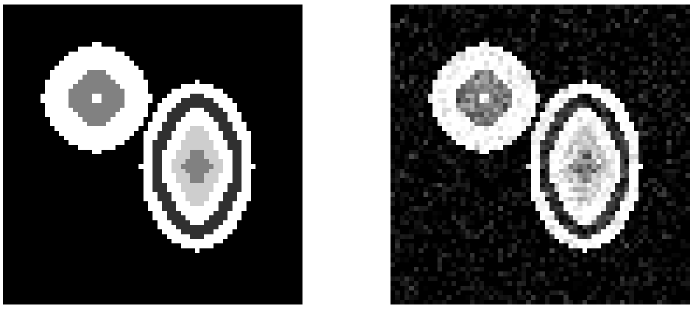
The trust-region subproblem (2.4.17)–(2.4.18) is solved only approximately. In fact, from our numerical experience one inner iteration seems sufficient for Algorithm 2.4.6 in practice. The outer loop is terminated once the residual norm $\|\nabla f_\gamma(u^k)\|$, see formula (2.3.5), has been reduced by a factor of 10^{-7}.

In step 5 of Algorithm 2.4.6, a R-regularized Newton system needs to be solved. In a denoising problem, i.e. when $K = I$, the linear system can be efficiently solved by sparse Cholesky factorization. For problems where K is a dense matrix or not even explicitly formulated as a matrix, we utilize the conjugate gradient method with residual tolerance 0.05. We remark that in our convergence analysis in section 2.4, step 5 is treated as exact equation solving. Nevertheless in the numerical realization, whenever the matrix $H_+^k + \beta^k R_{+,\varepsilon}^k$ is detected to be indefinite or (near-) singular, we immediately activate the sufficient condition for descent property (see Theorem 2.4.2), i.e. utilize step 7 of the algorithm.

All experiments were performed under MATLAB R2009b on a 2.66 GHz Intel Core Laptop with 4 GB RAM. The CPU time reported in the tables below is measured in seconds.

Test on "Two Circles" image

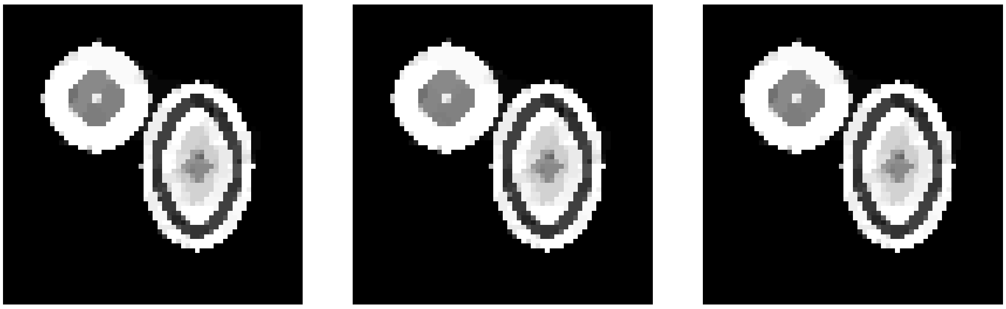
The 64-by-64 image "Two Circles" in [NNZC08] is used as our first test example, see Figure 2.1(a), in the context of a denoising problem.This image is corrupted by white Gaussian noise of zero mean and 0.1 standard deviation as shown in Figure 2.1(b). We choose the regularization parameters $\alpha = 2 \times 10^{-3}$ and $\mu = 0$ in the experiments.



(a) Original image. (b) Degraded image.

Figure 2.1: "Two Circles" image.

Dependence on initial guess. Three different choices of initial guesses are considered, namely the observed data, the zero vector, and a randomly chosen initial guess. The corresponding restored images are displayed in Figure 2.2, and the corresponding statistics are given in Table 2.1. We observe that the convergence behavior is stable with respect to the choice of the initial guess, in terms of both restoration quality and computational cost. Due to the nonconvex nature of the variational problem, our iterative algorithm is expected to terminate at a stationary point. In our experiments, the qualities of the obtained stationary points are almost equally good, in terms of objective value and PSNR (peak signal noise ratio), and all three restorations require about three seconds. In the sequel, we shall choose the observed data as our initial guess if not otherwise specified.



(a) $u^0 = z$. (b) $u^0 = 0$. (c) Randomly chosen u^0.

Figure 2.2: Dependence on initial guess.

Initial guess	Objective value	PSNR	CPU
$u^0 = z$	42.0354	30.2395	3.06
$u^0 = 0$	42.0385	30.2438	3.02
random u^0	42.0373	30.2048	3.19

Table 2.1: Dependence on initial guess.

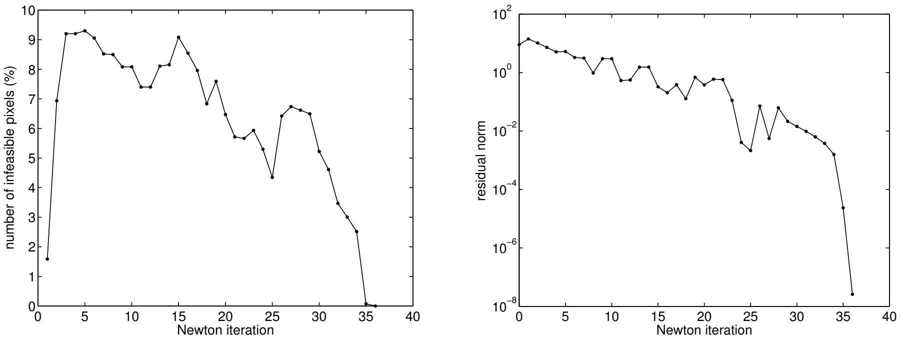
Dependence on Huber parameter γ. In the discrete variational model (2.3.4), the nondifferentiable TV^q-penalty term is locally smoothed by the Huber function φ_γ with Huber parameter γ. In Table 2.2, we show the results of numerical tests for four different choices of γ. It is observed that the convergence behavior of our algorithm is insensitive with respect to the choice of γ, once γ is sufficiently small. Clearly, with respect to γ there is a tradeoff between the convergence speed and the restoration quality. As γ goes to zero, one obtains higher restoration quality but at the same time the computational cost increases. From our experience, $\gamma = 0.1$ is practically a reasonable choice in general.

Huber parameter γ	1e1	1e0	1e-1	1e-2	1e-3
# Newton iter.	5	28	37	40	43
PSNR	25.3644	29.7011	30.12	30.1489	30.1489

Table 2.2: Dependence on Huber parameter γ.

Infeasible Newton technique. We note that in contrast to the primal-dual algorithm (for $q = 1$) in, e.g., [CGM99], our algorithm allows violations of the feasibility condition (2.4.12) during the Newton iterations. Yet towards the convergence of the algorithm we expect the feasibility condition (2.4.12) to hold true for (u^k, p^k), as established in the proof in Lemma 2.4.8. This is illustrated in Figure 2.3. In plot (a) the number of infeasible pixels $(i, j) \in \Omega$, where $|(\nabla u^k)_{ij}| \geq \gamma$ and $|(p^k)_{ij}||(\nabla u^k)_{ij}|^{1-q} \geq 1 + \epsilon_p$, is plotted for each Newton iteration. Here $\epsilon_p = 10^{-6}$ is introduced to compensate roundoff errors. In plot (b), the residual norm $\|\nabla f_\gamma(u^k)\|$ is shown for each Newton iteration. It is observed that the number of infeasible pixels decreases to zero as the algorithm converges.

Globalization by Wolfe-Powell line search. In Algorithm 2.4.6, after the search direction d^k is computed, the Wolfe-Powell line search is performed, where we aim to find an approximation of the solution to the one-dimensional problem $f_*^k := \min_{a^k > 0} f_\gamma(u^k + a^k d^k)$. Here we utilize an implementation according to Algorithm 3.5–3.6 in [NW06]. Essentially, we begin with an initial step size a^k equal to 1. If either this step size is acceptable or the interval $[0, 1]$ contains an acceptable step size (which we refer to as Case 1), we directly proceed to the *zoom* procedure [NW06], which successively reduces the size of the interval until an acceptable step size is found. Otherwise (which we refer to as Case 2), we keep increasing a^k until we find either an acceptable step size or a *solution interval* that contains the acceptable step size. Once the solution interval is found, we proceed to the zoom procedure as in Case 1. In Table 2.3 and Figure 2.4, we



(a) Number of infeasible pixels.

(b) Residual norm.

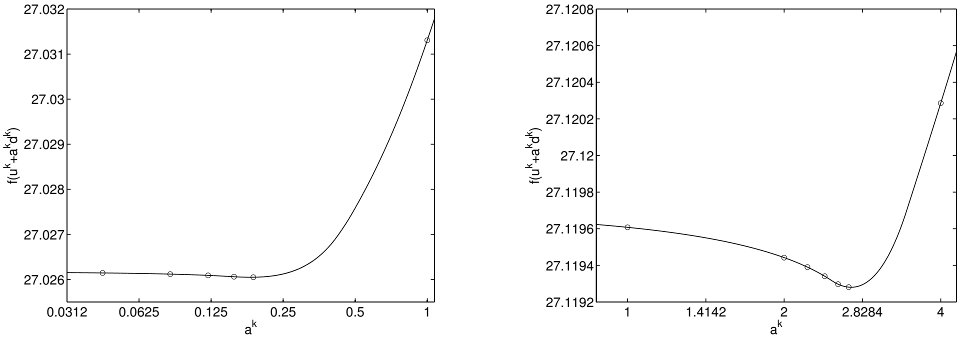
Figure 2.3: Infeasible Newton technique.

provide an example of the Wolfe-Powell line search for each of the two cases: zoom only (see the upper part of Table 2.3), and first increase a^k and then zoom (see the lower part of Table 2.3). We remark that backtracking-only line search rules, e.g. the backtracking Armijo rule (see, e.g., [NW06]), do not perform well in our context. A backtracking-only line search rule would terminate with $a^k = 1$ in the example for Case 2, which poorly resolves the line search problem and therefore causes more (outer) Newton iterations.

Case 1: zoom only							
a^k	1	0.04	0.084	0.122	0.156	0.188	
$f(u^k + a^k d^k) - f_*^k$	5.26e-3	9.57e-5	6.97e-5	4.13e-5	1.14e-5	4.06e-7	
Case 2: increase a^k and then zoom							
a^k	1	2	4	2.217	2.393	2.539	2.662
$f(u^k + a^k d^k) - f_*^k$	3.27e-4	1.62e-4	1.01e-3	1.1e-4	6.05e-5	1.7e-5	5.89e-7

Table 2.3: Wolfe-Powell line search history.

Comparison with existing algorithms. In Table 2.4, we compare Algorithm 2.4.6 with two existing algorithms, namely the BFGS quasi-Newton method (see e.g. [NW06]) and the lagged-diffusivity fixed-point method [VO96]. For a given tolerance with respect to the residual norm, we implement each candidate method with three different choices of the Huber parameter γ. The CPU time is reported in the corresponding entry. It is observed that our method always outperforms the other two methods, in particular when the problem becomes increasingly ill-conditioned as γ is reduced. We remark that the BFGS quasi-Newton method suffers from the strongly nonlinear nature of the underlying problem. The lagged-diffusivity fixed-point method performs reasonably well at early iterations, but becomes less competitive once higher accuracy is concerned.



(a) Case 1: zoom only.

(b) Case 2: increase a^k and then zoom.

Figure 2.4: Wolfe-Powell line search. In each figure, the solid line is a plot of the function $a^k \mapsto f_\gamma(u^k + a^k d^k)$, and the circled points are plots of the data in Table 2.3.

	BFGS		Fixed-point		Our method	
tolerance	1e-4	1e-7	1e-4	1e-7	1e-4	1e-7
γ=1e1	5.12	8.64	0.43	1.06	0.33	0.43
γ=1e0	51.21	70.91	3.98	12.54	1.86	2.68
γ=1e-1	>300	>300	4.7	20.02	2.44	3.07

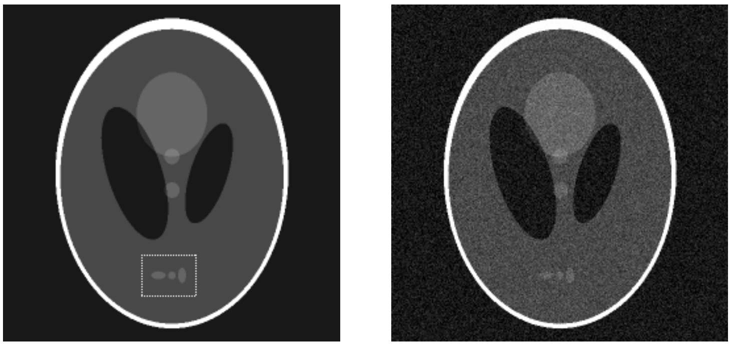
Table 2.4: Comparison with existing algorithms in terms of CPU time.

Test on "Shepp-Logan Phantom"

Our second testing image is the "Shepp-Logan Phantom" contaminated by white Gaussian noise of zero mean and 0.05 standard deviation.

Dependence on image resolution. Our algorithm is implemented to restore the "Shepp-Logan Phantom" images under different resolutions, namely 64-by-64, 128-by-128, and 256-by-256. The regularization parameters $\alpha = 4 \times 10^{-4}$ and $\mu = 0$ are fixed in all three restorations. The algorithm terminates after 62, 64, and 60 Newton iterations for restoring images under resolutions 64-by-64, 128-by-128, and 256-by-256, respectively. This indicates that our algorithm is stable with respect to the image resolution.

Performance of the TV^q-model for different q-values. We compare the performance of our TV^q-model for q =1, 0.75, 0.5, and 0.25 for denoising the 256-by-256 Shepp-Logan Phantom; see Figure 2.5. For each q, the parameter α is manually chosen in order to obtain the best PSNR value. The restored images $\widehat{u}_q$ are shown in Figure 2.6 for each q. It is observed from the rescaled zoom-in views that the TV^q-models provide better contrast in restoration as q becomes smaller. The performance of the TV^q-model for different q-choices is also compared quantitatively; see Table 2.5. The PSNR values of the restoration from the nonconvex TV^q-models (with $0 < q < 1$) are significantly higher than those from the TV-model (with $q = 1$). In addition, we measure the gradient sparsity by $|\mathcal{A}_\gamma(\widehat{u}_q)|/|\Omega|$, where $\mathcal{A}_\gamma(u) := \{(i,j) \in \Omega : |(\nabla u)_{ij}| \geq \gamma\}$. It is observed that in comparison with the conventional TV-model, the sparsity is well enhanced under the nonconvex TV^q-regularizations. Note that the gradient sparsity of the true image is 0.0503. Furthermore, we compare each solution of the TV^q-model, denoted by $\widehat{u}_q$, with the solution of the TV-model, denoted by $\widehat{u}_1$, by plugging both solutions into the objective of the nonconvex TV^q-problem. We find that $\widehat{u}_1$ is far from being optimal with respect to the objective value. This phenomenon is more distinct as q becomes smaller.



(a) Original image.

(b) Degraded image.

Figure 2.5: 256-by-256 Shepp-Logan Phantom. The dash-boxed region in (a) is zoomed in for restored images in forthcoming Figure 2.6.

q	PSNR	$\|\mathcal{A}_\gamma(\widehat{u}_q)\|/\|\Omega\|$	$f_{\gamma,q}(\widehat{u}_q)$	$f_{\gamma,q}(\widehat{u}_1)$
1	37.5709	0.196	-	-
0.75	41.0039	0.0578	134.3021	137.6719
0.5	41.0191	0.0531	116.6721	125.6655
0.25	39.9259	0.0503	113.9953	133.6758

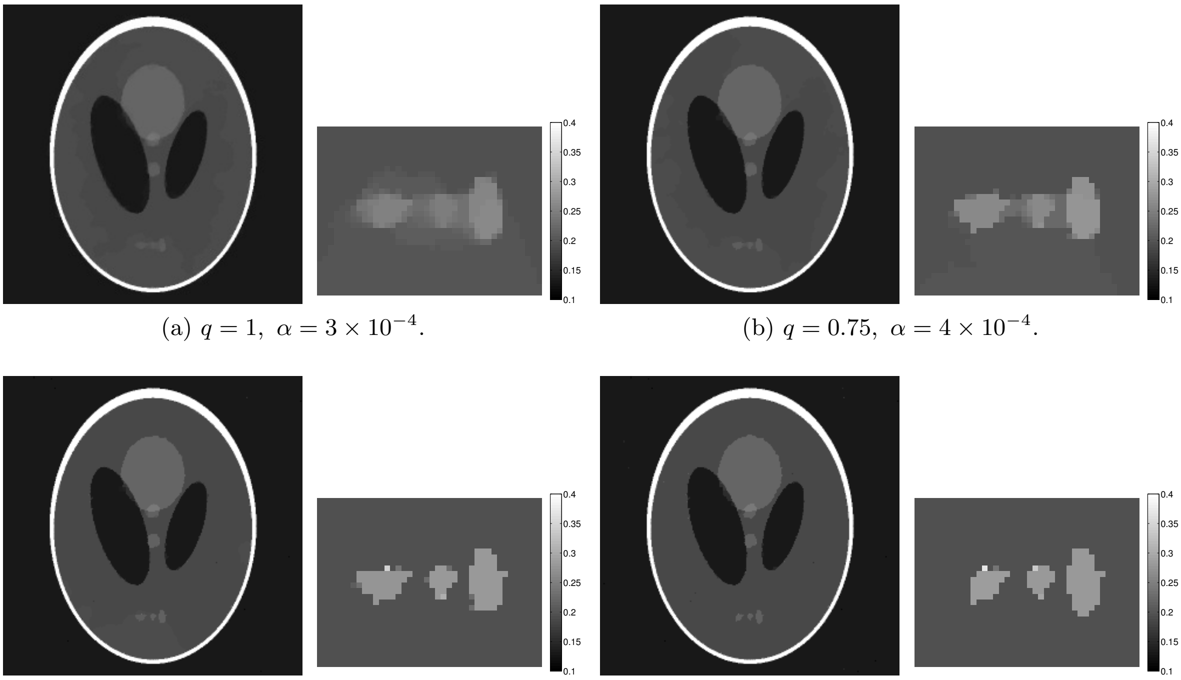
Table 2.5: Performances of TV^q-models.

Test on simultaneously blurred and noisy images

Now we apply our algorithm for simultaneously deblurring and denoising the text image "TV^q-model" (see Figure 2.7) and the image "Cameraman" (see Figure 2.8). For both images, the blurring kernel is chosen to be a two-dimensional truncated Gaussian kernel, i.e.

$$(Ku)_{ij} = \sum_{|i'|\leq 3,\ |j'|\leq 3} \exp\left(-\frac{|i'|^2+|j'|^2}{2|\sigma_K|^2}\right) u_{i-i',j-j'}.$$

with $\sigma_K = 1.5$. After blurring, white Gaussian noise of zero mean and 0.05 standard deviation is added. The restored images are shown in the corresponding figures. It is visually observed that the nonconvex TV^q-model promotes piecewise constant images in the restoration results. This is expected because $q \to 0$ results in the problem of minimizing the support of the image intensity.



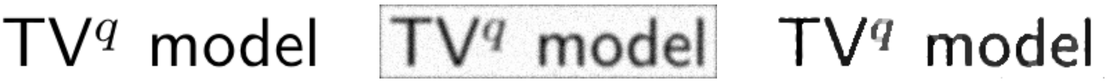
(a) $q = 1,\ \alpha = 3 \times 10^{-4}$.

(b) $q = 0.75,\ \alpha = 4 \times 10^{-4}$.

(c) $q = 0.5,\ \alpha = 6 \times 10^{-4}$.

(d) $q = 0.25,\ \alpha = 1.2 \times 10^{-3}$.

Figure 2.6: Restoration via TV^q-models. In each group, the left figure is the restored image $\widehat{u}_q$, and the right figure is the rescaled zoom-in of the restored image on the dash-boxed region of Figure 2.5(a).



(a) Original image.

(b) Observed image.

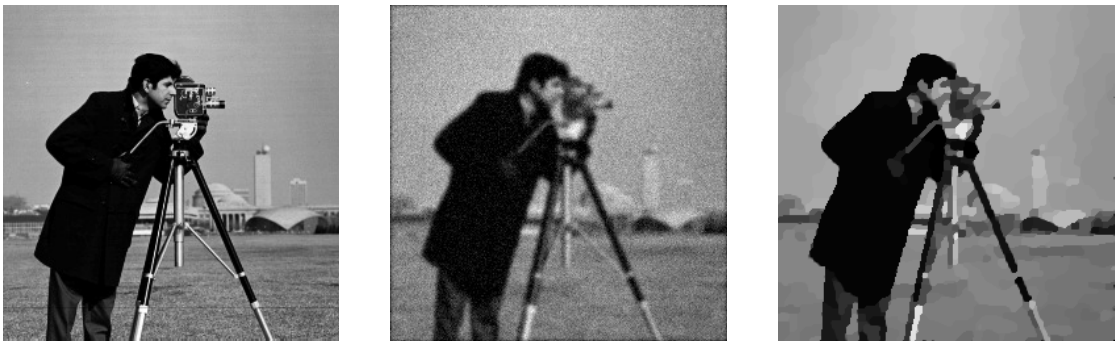
(c) Restored image.

Figure 2.7: "TV^q-model" text image: restoration with $\alpha = 5 \times 10^{-4}$.

Test on tomographic data

Our algorithm can be applied to restoring images from possibly noisy tomographic data. In Figure 2.9, the 64-by-64 Shepp-Logan Phantom is used as test example, see plot (a). The tomographic data, or the *sinogram*, of size 95-by-13 is obtained from applying the 2D Radon transform [KS01] along the angles 0, 12, 24, ..., 180 degrees. Then white Gaussian noise of zero mean and 0.05 standard deviation is added to the sinogram. The resulting observed data is shown in plot (b). Note that the matrix K is the discrete Radon transform of size 1235-by-4096, which indicates that the problem is highly underdetermined.

In our experiments, we consider three candidate methods, namely the filtered back-projection method (FBP) [KS01], the total-variation model (TV), and the TV^q-model, with $q = 0.75$, proposed in this work. The corresponding restored images are displayed in plots (c)–(e), and the comparisons of the three approaches in terms of PSNR and CPU time are given in Table 2.6.



(a) Original image.

(b) Observed image.

(c) Restored image.

Figure 2.8: “Cameraman” image: restoration with $\alpha = 2 \times 10^{-4}$.

FBP is implemented using the MATLAB routine `iradon`. For both TV- and TV^q-methods, we choose the regularization parameter $\alpha = 0.001$ and the initial guess $u^0 = 0$, and terminate the algorithm once the residual norm $\|\nabla f_\gamma(u)\|$ is reduced by a factor of 10^{-4}. It is observed that the computational cost of FBP is very low but the associated restoration quality is poor. The TV-method takes about 1.5 seconds and yields a much better restoration result, but some artifacts remain. Finally, the TV^q-method requires more CPU time than the other two methods (yet less than double the CPU time of the TV-method) but yields an almost perfect reconstruction.

	FBP	TV	TV^q
PSNR	16.5321	26.7974	37.3862
CPU	<0.01	1.56	2.64

Table 2.6: Comparison of restoration methods in terms of PSNR and CPU time.

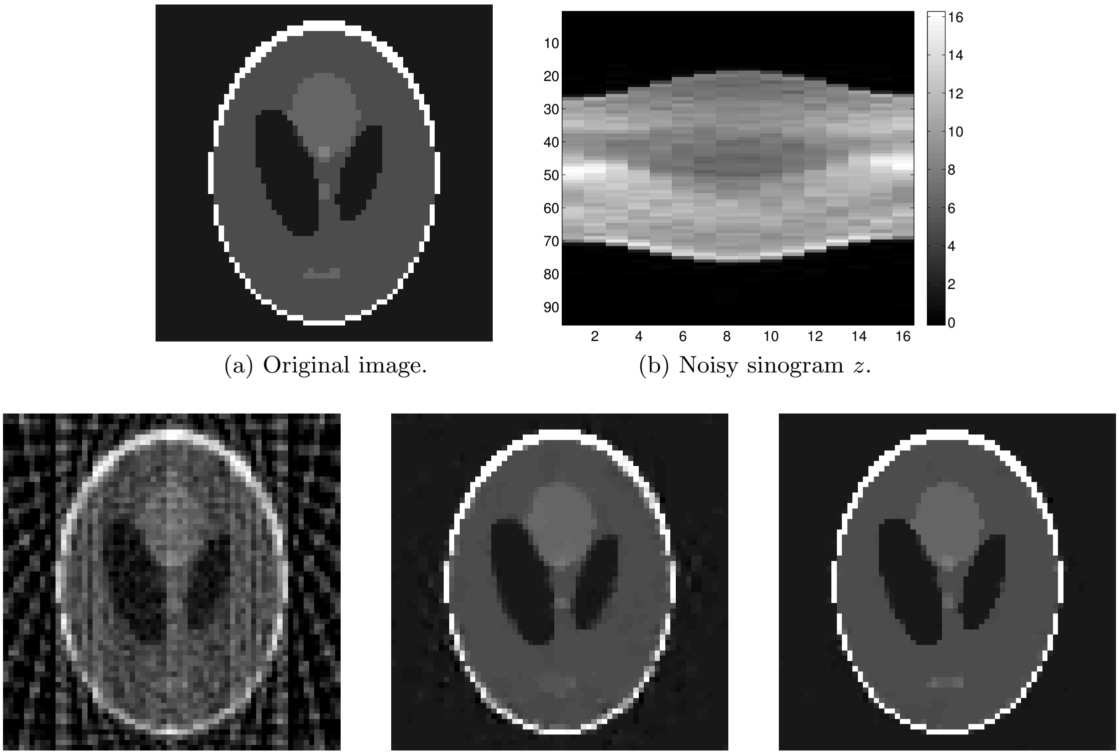
2.6 A smoothing scheme and the consistency result

In section 2.4, we have proposed an adaptively regularized Newton algorithm for solving (2.3.4) which is a Huberized version of the original problem (2.3.1). We further witness from the numerics in section 2.5 that, up to a reasonable choice of the Huber parameter γ, Algorithm 2.4.6 efficiently computes a numerical solution that is often satisfactory for practical concerns.

Nevertheless, we are intrigued by the question how to use the adaptively regularized Newton algorithm to track the solution of the original nonsmooth problem. Motivated by the recent findings in [Che12], here we provide a smoothing scheme with convergence analysis to accomplish this goal. It is substantiated by the convergence of the smoothing scheme below that the Huberization strategy provides a consistent approximation of the seemingly intractable nonsmooth problem (which is even non-Lipschitz in this case).

Algorithm 2.6.1 (Smoothing scheme)**.**

Require: parameters in Algorithm 2.4.6 and in addition $0 < \kappa_\gamma < 1,\ \rho_\gamma > 0$.



(a) Original image.

(b) Noisy sinogram z.

(c) Restored by FBP.

(d) Restored by TV.

(e) Restored by TV^q ($q = 0.75$).

Figure 2.9: Restoration from Radon transformed data.

```
Initialize the iterate $(u^0, p^0)$, the regularization scalar $\beta^0 \geq 0$, the trust-region radius $\sigma^0 > 0$,
and the Huber parameter $\gamma^0 > 0$. Set $k := 0$.
repeat
  Implement steps 3–25 in Algorithm 2.4.6 for the relaxed problem (2.3.4) with $\gamma := \gamma^k$.
  if $\|\nabla f_{\gamma^k}(u^k)\| > \rho_\gamma \gamma^k$ then
    Set $\gamma^{k+1} := \gamma^k$.
  else
    Set $\gamma^{k+1} := \kappa_\gamma \gamma^k$.
  end if
  Set $k := k + 1$.
until the stopping criterion for the smoothing scheme is fulfilled.
```

Lemma 2.6.2. *Let the sequence $\{u^k\}$ be generated by Algorithm 2.6.1. Then we have*

$$\lim_{k\to\infty} \gamma^k = 0 \quad \textit{and} \quad \liminf_{k\to\infty} \|\nabla f_{\gamma^k}(u^k)\| = 0.$$

Proof. Define the index set

$$\mathcal{K} := \{k : \gamma^{k+1} = \kappa_\gamma \gamma^k\}.$$

If $\mathcal{K}$ is finite, then there exists some $\bar{k}$ such that for all $k > \bar{k}$ we have $\gamma^k = \gamma^{\bar{k}}$ and $\|\nabla f_{\gamma^k}(u^k)\| \geq \rho_\gamma \gamma^{\bar{k}}$. This leads to a contradiction to the global convergence guaranteed by Theorem 2.4.7 that

$\lim_{k\to\infty} \|\nabla f_{\gamma^k}(u^k)\| = 0$. Thus, $\mathcal{K}$ must be infinite and $\lim_{k\to\infty} \gamma^k = 0$. Moreover, by ordering the indices in $\mathcal{K}$ as $k^1 < k^2 < k^3 < ...$, we have $\|\nabla f_{\gamma^{k^l}}(u^{k^l})\| \leq \rho_\gamma \gamma^{k^l} \to 0$ as $l \to \infty$. Hence, $\liminf_{k\to\infty} \|\nabla f_{\gamma^k}(u^k)\| = 0$. □

Theorem 2.6.3 (Consistency). *Assume that the sequence $\{u^k\}$ generated by Algorithm 2.6.1 is uniformly bounded. Then this sequence has an accumulation point u^* that satisfies the Euler-Lagrange equation (2.3.3).*

Proof. In view of the result in Lemma 2.6.2, there exists a subsequence of $\{u^k\}$, under the same notation, such that $\lim_{k\to\infty} \gamma^k = 0$ and $\lim_{k\to\infty} \|\nabla f_{\gamma^k}(u^k)\| = 0$. Let u^* be an accumulation point of the uniformly bounded sequence $\{u^k\}$. We show that u^* is a solution to (2.3.3). On the set $\{(i,j) \in \Omega : (\nabla u^*)_{ij} = 0\}$, the conclusion follows automatically. On the set $\{(i,j) \in \Omega : (\nabla u^*)_{ij} \neq 0\}$, we have $\max(|(\nabla u^k)_{ij}|, \gamma^k) \to |(\nabla u^*)_{ij}| > 0$ as $k \to \infty$. Therefore, it follows from

$$|(\nabla f(u^*))_{ij}| \leq |(\nabla f_{\gamma^k}(u^k))_{ij} - (\nabla f(u^*))_{ij}| + |(\nabla f_{\gamma^k}(u^k))_{ij}| \to 0, \tag{2.6.1}$$

that u^* satisfies (2.3.3). □

2.7 A note on TV^q-models in function space

Often one aims at studying the variational problem in its original function space setting. In our context, the infinite dimensional version associated with (2.3.1) reads

$$\inf_{u\in H_0^1(\Omega)} f(u) = \int_\Omega F(x,u,\nabla u)dx = \int_\Omega \left(\frac{\mu}{2}|\nabla u|^2 + \frac{\alpha}{q}|\nabla u|^q + \frac{\lambda}{2}|Ku-z|^2\right)dx, \tag{2.7.1}$$

where $\alpha > 0$, $0 < q < 1$, $0 < \mu \ll \alpha$, $z \in L^2(\Omega)$, $\lambda \in L^\infty(\Omega)$ such that $\lambda(x) > 0$ a.e. on Ω and $\int_\Omega \lambda(x)dx = \text{Area}(\Omega)$, and $K \in \mathcal{L}(L^2(\Omega))$, i.e. it is a linear and continuous operator from $L^2(\Omega)$ to $L^2(\Omega)$, such that $K\chi_\Omega \neq 0$.

Obviously, f is coercive, i.e. $f(u) \to \infty$ as $\|u\|_{H_0^1(\Omega)} \to \infty$. Note that the integrand $F(x,u,\xi)$ is nonconvex in ξ. It is known from Theorem 2.1.3 in [AK02] that f is weakly lower semicontinuous on $H_0^1(\Omega)$ if and only if F is convex in ξ. As a consequence, $f : H_0^1(\Omega) \to \mathbb{R}$ in (2.7.1) is not weakly lower semicontinuous, a usual prerequisite for proving existence of minimizers. Hence, the direct methods of the calculus of variations cannot be applied here.

Nevertheless, there exists a minimizer for a relaxed version of the problem (2.7.1). For this purpose, we construct a relaxed functional by taking the bipolar [ET99] of $F(x,u,\xi)$ with respect to ξ, i.e.

$$\bar{F}(x,u,\xi) := F^{**}(x,u,\xi) = \begin{cases} (\alpha s_*^{q-1} + \mu s_*)|\xi| + \frac{\lambda}{2}|Ku-z|^2, & \text{if } |\xi| < s_*, \\ F(x,u,\xi), & \text{if } |\xi| \geq s_*, \end{cases}$$

where the convexity threshold s_* is given by

$$s_*(q,\mu,\alpha) := \left(\frac{\alpha(1/q-1)}{\mu/2}\right)^{1/(2-q)}.$$

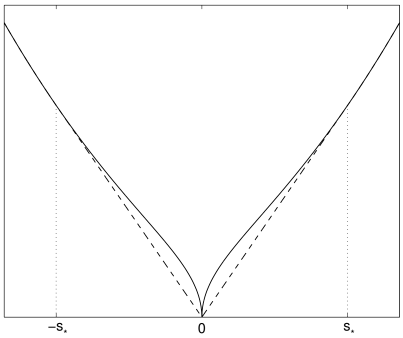


Figure 2.10: The function $s \mapsto \frac{\mu}{2}|s|^2 + \frac{\alpha}{q}|s|^q$ (in solid line), and its envelope (in dashed line).

We define $\bar{f}(u) := \int_\Omega \bar{F}(x, u, \nabla u)dx$. It turns out that $\bar{f}$ represents the weakly lower semicontinuous envelope of $f(u)$ under the weak $H_0^1(\Omega)$-topology (see pp. 34 in [DM93]), i.e.

$$\bar{f}(u) = \sup\{\widetilde{f}(u) : \widetilde{f}(u) \leq f(u) \; \forall u \in H_0^1(\Omega), \; \widetilde{f} \text{ is weakly lower semicontinuous on } H_0^1(\Omega)\}.$$

Concerning the existence of minimizers in $H_0^1(\Omega)$ for $\bar{f}$ and their relations to f, we state the following two theorems, which can be found in [AK02]; see Theorem 2.1.5 and Theorem 2.1.6 in this reference.

Theorem 2.7.1 (Characterization). *The relaxed functional $\bar{f}$ is characterized by the following properties:*

1. *For every sequence $\{u^k\}$ that weakly converges to u in $H_0^1(\Omega)$, we have $\bar{f}(u) \leq \liminf f(u^k)$.*
2. *For every $u \in H_0^1(\Omega)$, there exists a sequence $\{u^k\}$ that weakly converges to u in $H_0^1(\Omega)$ and $\bar{f}(u) \geq \limsup f(u^k)$.*

Theorem 2.7.2 (Main properties). *Suppose $f : H_0^1(\Omega) \to \mathbb{R}$ is coercive. Then the following properties hold:*

1. *$\bar{f}$ is coercive and weakly lower semicontinuous on $H_0^1(\Omega)$.*
2. *$\bar{f}$ has a minimizer in $H_0^1(\Omega)$.*
3. $\min_{u \in H_0^1(\Omega)} \bar{f}(u) = \inf_{u \in H_0^1(\Omega)} f(u)$.
4. *Every accumulation point of an infimizing sequence for f is a minimizer for $\bar{f}$ under the weak $H_0^1(\Omega)$-topology.*
5. *Every minimizer for $\bar{f}$ is the limit of an infimizing sequence for f under the weak $H_0^1(\Omega)$-topology.*

In a nutshell, we associate the original nonconvex problem, for which no minimizer may exist, with a relaxed problem, which ensures the existence of a minimizer. However, the minimizer of the relaxed problem may be far from optimal for the original problem with respect to the objective value. This is illustrated by the following example; see pp. 36 in [AK02] for a related example. Note that this example shares the nonconvexity in the ξ-variable with our TV^q-model, but otherwise has a different structure in the term involving the derivative.

Example 2.7.3. Let $F(x, u, \xi) := u^2 + (|\xi| - 1)^2$. The Bolza problem is

$$\inf\{f(u) := \int_0^1 \left((|u'| - 1)^2 + u^2\right) dx : u \in H_0^1(0,1)\}.$$

The integrand $F(x, u, \xi)$ is nonconvex in ξ. We claim that $\inf f = 0$. Indeed, consider the sequence $\{u^k\}$ defined by

$$u^k(x) = \begin{cases} x - \frac{l}{k} & \text{if } x \in \left(\frac{l}{k}, \frac{2l+1}{2k}\right) \\ -x + \frac{l+1}{k} & \text{if } x \in \left(\frac{2l+1}{2k}, \frac{l+1}{k}\right) \end{cases}, \text{ for } l = 0, 1, 2, ..., n-1.$$

Then $u^k \in W_0^{1,\infty}(0,1)$ such that $0 \leq u^k(x) \leq \frac{1}{2k}, \forall x \in (0,1)$, and $|(u^k)'(x)| = 1$ a.e. in $(0,1)$. Therefore, we have $0 \leq \inf_u f(u) \leq f(u^k) \leq \frac{1}{4k^2}$. Thus the claim is verified. However, there exists no function $u \in H_0^1(0,1)$ such that $f(u) = 0$. Hence there exists no solution to the Bolza problem.

Nevertheless, the Bolza problem can be relaxed, using the weakly lower semicontinuous envelope of f, as follows:

$$\min\{\bar{f}(u) := \int_0^1 \left((\max(|u'| - 1, 0))^2 + u^2\right) dx : u \in H_0^1(0,1)\}.$$

The relaxed problem admits a unique solution $u^* = 0$. Obviously the set $\{x \in (0,1) : |(u^*)'(x)| < 1\}$ is of positive Lebesgue measure; otherwise u^* would be a minimizer for $\inf_u f(u)$. Finally, we notice that $f(u^*) = 1$. This indicates that u^* is far from optimal for the original problem.

2.8 Beyond TV^q — variational models with concave sparsity-promoting priors

Our methodology developed so far in this chapter focuses on the TV^q models for image restoration. It is no surprise that this methodology fits into a wider scope. The generalization of the TV^q-models in this section is twofold. First, we free the prior term from particular parameterizations to a general class of concave priors, which includes the ℓ^q-norm ($0 < q < 1$) as a special case. Secondly, while the TV^q-models promote piecewise constant images which are gradient-sparse, in some other applications we are seeking sparse solutions under other specified transforms. Following a similar route as for the TV^q models, we also devise a superlinearly convergent Newton-type method for the generalized variational model (under Huberization). Our numerical experiments will demonstrate the applications of the generalized methodology in dictionary-based image denoising, support vector machines, and optimal control of partial differential equations.

2.8.1 Variational models with concave priors

We consider the following general variational model:

$$\min_{u \in \mathbb{R}^{|\Omega_u|}} f(u) = \Theta(u) + \alpha\Psi(u), \tag{2.8.1}$$

where Ω_u denotes the multidimensional index set for u. We assume that the fidelity term $\Theta : \mathbb{R}^{|\Omega_u|} \to \mathbb{R}$ is a coercive and strictly convex C^2-function. Thus, the Hessian $\nabla^2\Theta(\cdot)$ exists and is positive definite everywhere in $\mathbb{R}^{|\Omega_u|}$.

The prior term Ψ under consideration is of the form

$$\Psi(u) = \sum_{j \in \Omega_p} \psi(|(Gu)_j|),$$

where $G : \mathbb{R}^{|\Omega_u|} \to \mathbb{R}^{|\Omega_p|}$ is a bounded linear operator and Ω_p is the multidimensional index set for a transformed vector Gu. The scalar function $\psi : [0, \infty) \to [0, \infty)$ is supposed to satisfy the following hypotheses:

1. (continuity) ψ is continuous on $[0, \infty)$.

2. (regularity) ψ is C^2 on $(0, \infty)$.

3. (mononicity) ψ is strictly increasing on $[0, \infty)$.

4. (concavity) ψ is concave on $[0, \infty)$.

The motivation for a concave prior ψ is to sparsify the solution u under a certain transform G; see e.g. [Nik05]. Typical choices for G include the identity [FL01], the gradient operator [HW13], or some overcomplete dictionary [KP13]. In particular, we are interested in those situations where $\psi(|\cdot|)$ is non-smooth or even non-Lipschitz at 0. Particular examples for ψ, which have been considered in either a statistical or variational framework, are specified below.

Example 2.8.1 (Concave priors).

- Bridge prior [KF00, HHM08]: $\psi(s) = s^q/q$, $0 < q < 1$.

- Fraction prior [GR92]: $\psi(s) = qs/(1 + qs)$, $q > 0$.

- Logarithmic prior [Nik05]: $\psi(s) = \log(1 + qs)$, $q > 0$.

The proof of existence of a solution for (2.8.1) is straightforward due to the fact that the objective f is continuous, coercive, and bounded from below. In order to characterize a stationary point for (2.8.1), we introduce an auxiliary variable $p \in \mathbb{R}^{|\Omega_p|}$ and derive the Euler-Lagrange equation as follows.

Theorem 2.8.2 (Necessary optimality condition). *For any global minimizer of (2.8.1) there exists some* $p \in \mathbb{R}^{|\Omega_p|}$ *such that*

$$\begin{cases} \nabla\Theta(u) + \alpha G^\top p = 0, \\ \varphi(|(Gu)_j|)p_j = (Gu)_j, \quad \text{for all } j \in \Omega_p \text{ with } (Gu)_j \neq 0, \end{cases} \tag{2.8.2}$$

where $\varphi(s) := s/\psi'(s)$ *for any* $s \in (0, \infty)$.

Note that since (2.8.1) is a nonconvex minimization problem, in general there exist more than one stationary point satisfying the Euler-Lagrange system (2.8.2).

To handle the non-smoothness (or even non-Lipschitz continuity) of $\psi(|\cdot|)$ numerically, we introduce a Huber-type local smoothing [Hub64] by defining

$$\psi_\gamma(s) = \begin{cases} \psi(s) - \left(\psi(\gamma) - \dfrac{\gamma\psi'(\gamma)}{2}\right), & \text{if } |s| \geq \gamma, \\ \dfrac{\psi'(\gamma)}{2\gamma}s^2, & \text{if } |s| < \gamma, \end{cases}$$

where $\gamma > 0$ is the associated Huber parameter. Then we replace ψ in (2.8.1) by the C^1 function ψ_γ and formulate the *Huberized* variational model as:

$$\min_{u \in \mathbb{R}^{|\Omega_u|}} f_\gamma(u) = \Theta(u) + \alpha \sum_{j \in \Omega_p} \psi_\gamma(|(Gu)_j|). \tag{2.8.3}$$

The corresponding Euler-Lagrange equation for (2.8.3), which we call the Huberized Euler-Lagrange equation, is given by

$$\nabla f_\gamma(u) = \nabla\Theta(u) + \alpha G^\top(\varphi(\max(|Gu|, \gamma))^{-1}Gu) = 0, \tag{2.8.4}$$

or equivalently posed with an auxiliary variable p as follows:

$$\operatorname{res}(u, p; \gamma) := \begin{bmatrix} \nabla\Theta(u) + \alpha G^\top p \\ \varphi(\max(|Gu|, \gamma))p - Gu \end{bmatrix} = 0. \tag{2.8.5}$$

Since the argument of φ is bounded below by the positive number γ, the quantity $\varphi(\cdot)$ is well defined. Furthermore, φ satisfies the following properties: (1) φ is continuously differentiable on $(0, \infty)$; (2) $\varphi'(s) \geq \psi'(\gamma)^{-1} > 0$ for any $s \in [\gamma, \infty)$. Consequently, by the composition rule of semismooth functions Theorem 19 in [Fis97], the residual function $\operatorname{res}(\cdot, \cdot; \gamma)$ is semismooth at any $(u, p) \in \mathbb{R}^{|\Omega_u|} \times \mathbb{R}^{|\Omega_p|}$. This allows us to apply the semismooth Newton method to (2.8.5) as we shall see in the next subsection.

2.8.2 A superlinearly convergent regularized Newton scheme

In this subsection, we propose a tailored approach for finding a stationary point for (2.8.3). We start by investigating a structured regularization scheme in the semismooth Newton method.

R-regularized Newton scheme

Let (u^k, p^k) be the current iterate and the active set characteristic $\chi_{\mathcal{A}^k} \in \mathbb{R}^{|\Omega_p|}$ be defined as

$$(\chi_{\mathcal{A}^k})_j = \begin{cases} 1, & \text{if } |(Gu^k)_j| \geq \gamma, \\ 0, & \text{if } |(Gu^k)_j| < \gamma. \end{cases}$$

The set $\mathcal{A}^k := \{j \in \Omega_p : (Gu^k)_j| \geq \gamma\}$ is referred to as the active set. In view of the max-function, we shall apply the semismooth Newton method to (2.8.5). This leads us to the following linear system

$$\begin{bmatrix} \nabla^2\Theta(u^k) & \alpha G^\top \\ -\mathrm{diag}\left(1 - \chi_{\mathcal{A}^k} p^k \dfrac{\varphi'(m^k)Gu^k}{m^k}\right) G & \mathrm{diag}(\varphi(m^k)) \end{bmatrix} \begin{bmatrix} \delta u^k \\ \delta p^k \end{bmatrix} = \begin{bmatrix} -\nabla\Theta(u^k) - \alpha G^\top p^k \\ -\varphi(m^k)p^k + Gu^k \end{bmatrix},$$

with

$$m^k := \max(|Gu^k|, \gamma).$$

After eliminating δp^k, we are left with

$$H^k \delta u^k = -g^k,$$

where

$$\begin{aligned} H^k &= H(u^k, \chi_{\mathcal{A}^k} p^k) \\ &= \nabla^2\Theta(u^k) + \alpha G^\top \mathrm{diag}\left(\varphi(m^k)^{-1}\left(1 - \chi_{\mathcal{A}^k} p^k \frac{\varphi'(m^k)(Gu^k)}{m^k}\right)\right) G, \qquad (2.8.6) \\ g^k &= \nabla f_\gamma(u^k) = \nabla\Theta(u^k) + \alpha G^\top(\varphi(m^k)^{-1} Gu^k). \end{aligned}$$

Based on an observation of the structure of the Hessian matrix H^k (see equation (2.8.7) below), we are motivated to define the *R-regularization* of H^k at $(u^k, \chi_{\mathcal{A}^k} p^k)$ as

$$R^k = R(u^k, \chi_{\mathcal{A}^k} p^k) = \alpha G^\top \mathrm{diag}\left(\chi_{\mathcal{A}^k} p^k \frac{\varphi'(m^k)(Gu^k)}{\varphi(m^k)m^k}\right) G.$$

Then the resulting R-regularized Newton scheme arises as

$$(H^k + \beta R^k)\delta u^k = -g^k.$$

In particular, if we take $\beta = 1$, then the R-regularized Newton scheme becomes

$$(H^k + R^k)\delta u^k = \left(\nabla^2\Theta(u^k) + \alpha G^\top \mathrm{diag}(\varphi(m^k)^{-1})G\right)\delta u^k = -g^k. \qquad (2.8.7)$$

Note that the fully R-regularized Hessian $H^k + R^k$ is strictly positive definite and thus guarantees a descent direction δu^k for f_γ at u^k.

Infeasible Newton technique

In order to ensure fast local convergence of the overall Newton scheme, we introduce several modifications in the construction of H^k and R^k.

We start by replacing $\chi_{\mathcal{A}^k} p^k$ by $\widetilde{p}^k$ in formula (2.8.6), where

$$\widetilde{p}^k := \frac{\chi_{\mathcal{A}^k}(m^k/\varphi(m^k))p^k}{\max(m^k/\varphi(m^k), |p^k|)}.$$

This choice of $\widetilde{p}^k$ satisfies the feasibility condition

$$|(\widetilde{p}^k)_j| \leq |(Gu^k)_j|/\varphi(|(Gu^k)_j|),$$

on the index subset $\{j \in \Omega_p : |(Gu^k)_j| \geq \gamma\}$. As a consequence, the modified Hessian $\widetilde{H}^k$ appears as

$$\widetilde{H}^k = H(u^k, \widetilde{p}^k) = \nabla^2\Theta(u^k) + \alpha G^\top \mathrm{diag}\left(\varphi(m^k)^{-1}\left(1 - \widetilde{p}^k \frac{\varphi'(m^k)(Gu^k)}{m^k}\right)\right) G.$$

One of our motivations for such a replacement is that the sequence $(\widetilde{p}^k)$ will be uniformly bounded provided that the sequence (u^k) is uniformly bounded, which will be useful in the derivation of global convergence; see Theorem 2.8.9 below. In addition, the Hessian modification becomes asymptotically invariant, as shown in the following lemma.

Lemma 2.8.3. *Assume that* $\lim_{k\to\infty}(u^k, p^k) = (u^*, p^*)$ *with the limiting pair* (u^*, p^*) *satisfying the Euler-Lagrange equation (2.8.5). Then we have*

$$\lim_{k\to\infty} \|\widetilde{H}^k - H^k\| = 0.$$

Proof. Based on the structures of $\widetilde{H}^k$ and H^k, it suffices to show $\lim_{k\to\infty} \|\widetilde{p}^k - \chi_{\mathcal{A}^k}p^k\| = 0$. Given the assumption, we have for all $j \in \Omega_p$ that $|p^*| = |Gu^*|/\varphi(\max(|Gu^*|, \gamma))$ and therefore

$$\begin{aligned}|\widetilde{p}^k - \chi_{\mathcal{A}^k}p^k| &\leq |p^k| \left| \frac{m^k/\varphi(m^k)}{\max(m^k/\varphi(m^k), |p^k|)} - 1 \right| \\ &\to |p^*| \left| \frac{\max(|Gu^*|, \gamma)/\varphi(\max(|Gu^*|, \gamma))}{\max(\max(|Gu^*|, \gamma)/\varphi(\max(|Gu^*|, \gamma)), |p^*|)} - 1 \right| = 0\end{aligned}$$

as $k \to \infty$. Thus the conclusion follows. □

Besides the Hessian modification, we correspondingly define the modified R-regularization by

$$\widetilde{R}^k = R(u^k, \widetilde{p}^k) = \alpha G^\top \mathrm{diag}\left(\widetilde{p}^k \frac{\varphi'(m^k)(Gu^k)}{\varphi(m^k)m^k}\right) G + \varepsilon I, \tag{2.8.8}$$

with an arbitrarily fixed parameter $0 < \varepsilon \ll \alpha$.

Lemma 2.8.4. *Let the assumptions of Lemma 2.8.3 hold true. Then we have* $\lambda_{\min}(\widetilde{R}^k) \geq \varepsilon/2$ *for all sufficiently large* k.

Proof. As the conclusion is trivial given $G = 0$, we proceed with any given $G \neq 0$. It follows from the assumption that

$$\begin{aligned}
&\min\left\{\left(\widetilde{p}^k\frac{\varphi'(m^k)(Gu^k)}{\varphi(m^k)m^k}\right)_j : j \in \Omega_p\right\}\\
=&\min\left\{\left(\frac{\varphi'(m^k)(Gu^k)\chi_{\mathcal{A}^k}(m^k/\varphi(m^k))p^k}{\varphi(m^k)m^k\max(m^k/\varphi(m^k),|p^k|)}\right)_j : j \in \Omega_p\right\}\\
\geq&\min\left(\left\{\left(\frac{\varphi'(m^k)(Gu^k)(m^k/\varphi(m^k))p^k}{\varphi(m^k)m^k\max(m^k/\varphi(m^k),|p^k|)}\right)_j : j \in \Omega_p\right\}\cup\{0\}\right)\\
\xrightarrow{k\to\infty}&\min\left(\left\{\left(\frac{\varphi'(\max(|Gu^*|,\gamma))(Gu^*)^2}{\max(|Gu^*|,\gamma)(\varphi(\max(|Gu^*|,\gamma)))^2}\right)_j : j \in \Omega_p\right\}\cup\{0\}\right)\\
\geq&0.
\end{aligned}$$

Therefore, we have for all sufficiently large k that

$$\min\left\{\widetilde{p}^k\frac{\varphi'(m^k)(Gu^k)}{\varphi(m^k)m^k} : j \in \Omega_p\right\} \geq -\frac{\varepsilon}{2\|G\|^2},$$

and

$$v^\top \widetilde{R}^k v \geq -\frac{\varepsilon}{2\|G\|^2}\|Gv\|^2 + \varepsilon\|v\|^2 \geq \frac{\varepsilon}{2}\|v\|^2,$$

for any vector $v \in \mathbb{R}^{|\Omega_p|}$. Thus we conclude that $\lambda_{\min}(\widetilde{R}^k) \geq \varepsilon/2$. □

The ε-term in (2.8.8) is important as indicated by Lemma 2.8.4 since it guarantees $\widetilde{R}^k$ to be strictly positive definite when the iterate is sufficiently close to a solution. However, note that ε can be arbitrarily small and therefore $\widetilde{R}^k$ is allowed to have nonpositive eigenvalues during the Newton iterations. In fact, choosing a large ε that dominates the R-regularization term is not desirable in the numerical implementation.

Thus far, we arrive at the overall modified R-regularized Newton scheme

$$(\widetilde{H}^k + \beta\widetilde{R}^k)\delta u^k = -g^k. \tag{2.8.9}$$

The fully R-regularized scheme, i.e. with $\beta = 1$, generates a descent direction satisfying the estimate in the following theorem.

Theorem 2.8.5 (Sufficient condition for descent property)**.** *Assume that the sequence (u^k) is uniformly bounded and contained in a compact subset E in $\mathbb{R}^{|\Omega_u|}$. Then the solution δu^k of (2.8.9) with $\beta = 1$ is a descent direction satisfying*

$$-\frac{(g^k)^\top\delta u^k}{\|g^k\|\|\delta u^k\|} \geq \frac{C_l}{C_u + \alpha\lambda_{\max}(G^\top G)/\varphi(\gamma)} =: \bar{\epsilon}_d,$$

where $0 < C_l \leq C_u$ are two constants depending on Θ and E.

Proof. Let $S = \{v \in \mathbb{R}^{|\Omega_u|} : \|v\| = 1\}$ denote the unit sphere. Due to the compactness of $E \times S$ and the continuity of the functional $(u, v) \mapsto v^\top \nabla^2 \Theta(u) v$, the problem

$$C_l := \inf_{(u,v)\in E\times S} v^\top \nabla^2 \Theta(u) v$$

attains the infimum C_l for some $(\underline{u}, \underline{v}) \in E \times S$. Note that $C_l > 0$, since otherwise our assumption that Θ is a strictly convex C^2 function would be violated.

Analogously, there exists a constant C_u such that

$$C_u := \sup_{(u,v)\in E\times S} v^\top \nabla^2 \Theta(u) v.$$

Obviously, we have $C_u \geq C_l$. Then it follows that

$$\begin{aligned} &-\frac{(g^k)^\top \delta u^k}{\|g^k\|\|\delta u^k\|} \geq \frac{\lambda_{\min}(\widetilde{H}^k + \widetilde{R}^k)}{\lambda_{\max}(\widetilde{H}^k + \widetilde{R}^k)} \\ \geq &\frac{\lambda_{\min}(\nabla^2\Theta(u^k))}{\lambda_{\max}(\nabla^2\Theta(u^k)) + \lambda_{\max}(\alpha G^\top \mathrm{diag}(\varphi(m^k)^{-1})G)} \\ \geq &\frac{C_l}{C_u + \alpha\lambda_{\max}(G^\top G)/\varphi(\gamma)}. \end{aligned}$$

For the last inequality, we have used the fact that φ is monotonically increasing on $[\gamma, \infty)$. □

A superlinearly convergent algorithm

Analogous to Algorithm 2.4.6, here we also devise a superlinearly convergent algorithm for (2.8.1). According to Theorem 2.8.5, the R-regularized Newton scheme (2.8.9) with $\beta = 1$ provides a descent direction. However, a constant R-regularization (with $\beta = 1$), which is equivalent to a fixed-point approach, is known to be only linearly convergent [VO96, CM99, NC07].

Ideally, we would like to utilize a sufficient R-regularization when the objective is nonconvex (or the Hessian possesses negative eigenvalues) at the current iterate. As the iterative scheme proceeds, the iterate may eventually be contained in a neighborhood of some local minimizer satisfying some type of a second-order sufficient optimality condition, such that all (generalized) Hessians of the objective are positive definite within that neighborhood. Under such circumstances, we would rather utilize the true Hessian without any R-regularization in the Newton scheme, as it leads to local superlinear convergence.

In order to achieve these goals, the weight of the R-regularization β will be adjusted under a trust-region framework. Define the local quadratic model of f_γ at the current iterate u^k as

$$h^k(d) := f_\gamma(u^k) + (g^k)^\top d + \frac{1}{2} d^\top \widetilde{H}^k d.$$

Consider now the minimization of $h^k(\cdot)$ subject to a trust-region constraint, i.e.

$$\begin{aligned} &\text{minimize} && h^k(d), && (2.8.10) \\ &\text{subject to} && d \in \mathbb{R}^{|\Omega_u|},\ d^\top \widetilde{R}^k d \leq (\sigma^k)^2, && (2.8.11) \end{aligned}$$

where $\sigma > 0$ is the trust-region radius.

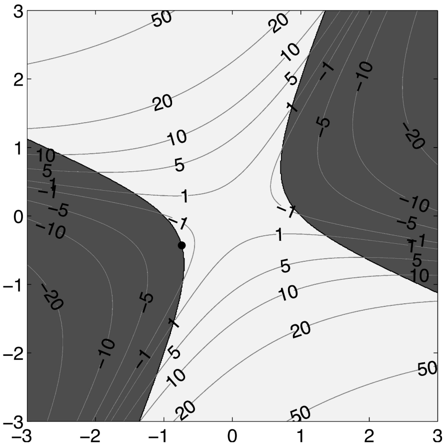


Figure 2.11: Illustration of a two-dimensional trust-region subproblem (2.8.10)–(2.8.11). The objective function is plotted with contour lines. The feasible region is colored in light gray (contrary to dark gray). The global minimizer $(-0.748, -0.403)$ is marked by the solid dot.

Note that the matrix $\widetilde{H}^k$ may be indefinite due to the nonconvexity of f_γ. Furthermore, as pointed out after Lemma 2.8.4, $\widetilde{R}^k$ is allowed to have more than one nonpositive eigenvalues. Thus, the feasible region induced by (2.8.11) may be nonconvex and unbounded; see Figure 2.11 for an illustration in two dimensions. This is significantly different from the settings in classical trust-region methods [DS96, CGT00] where $\widetilde{R}^k$ is positive definite and induces a convex, closed and bounded feasible region. Remarkably, $\widetilde{H}^k$ and $\widetilde{R}^k$ enjoy a special interplay as indicated in the following lemma.

Lemma 2.8.6. *The matrix $\widetilde{H}^k$ is positive definite on the subset $\{d \in \mathbb{R}^{|\Omega_u|} : d^\top \widetilde{R}^k d \le 0\}$.*

Proof. See Lemma 2.4.3. □

Such an interplay between $\widetilde{H}^k$ and $\widetilde{R}^k$ leads us to the existence as well as the characterization of a global minimizer for the trust-region subproblem (2.8.10)–(2.8.11), as stated in the following theorem.

Theorem 2.8.7. *There exists a global minimizer d_* for (2.8.10)–(2.8.11). Moreover, the necessary and sufficient condition for d_* being optimal is that there exists $\beta_* \ge 0$ such that*

$$(\widetilde{H}^k + \beta_* \widetilde{R}^k) d_* = -g, \tag{2.8.12}$$

$$\beta_* - \max\Big(\beta_* + c^{-1}(d_*^\top \widetilde{R}^k d_* - (\sigma^k)^2), 0\Big) = 0, \tag{2.8.13}$$

$$\widetilde{H}^k + \beta_* \widetilde{R}^k \succeq 0, \tag{2.8.14}$$

for an arbitrarily fixed scalar $c > 0$.

Proof. See Theorems 2.4.4 and 2.4.5. □

In particular, the complementarity equation (2.8.13) in Theorem 2.8.7 provides us a natural fixed-point formula for updating the weight β. Now we are in a position to present our superlinearly convergent R-regularized Newton scheme for (2.8.1).

Algorithm 2.8.8 (Superlinearly convergent R-regularized Newton scheme)**.**

```
Require: parameters $c > 0$, $0 < \rho_1 \leq \rho_2 < 1$, $0 < \kappa_1 < 1 < \kappa_2$, $0 < \varepsilon \ll \alpha$, $0 < \epsilon_d \leq \bar{\epsilon}_d$, $0 < \tau_1 < 1/2$, $\tau_1 < \tau_2 < 1$.
 1: Initialize the iterate $(u^0, p^0)$, the regularization weight $\beta^0 \geq 0$, and the trust-region radius $\sigma^0 > 0$. Set $k := 0$.
 2: repeat {outer loop}
 3:   Generate $\widetilde{H}^k$, $\widetilde{R}^k$, and $g^k$ at the current iterate $(u^k, p^k)$.
 4:   repeat {inner loop}
 5:     Solve the linear equation $(\widetilde{H}^k + \beta^k \widetilde{R}^k) d^k = -g^k$ for $d^k$.
 6:     if the matrix $\widetilde{H}^k + \beta^k \widetilde{R}^k$ is singular or $-(g^k)^\top d^k / (\|g^k\| \|d^k\|) < \epsilon_d$ then
 7:       Set $\beta^k := 1$, and return to Step 5.
 8:     end if
 9:     if $\beta^k = 1$ and $(d^k)^\top \widetilde{R}^k d^k > (\sigma^k)^2$ then
10:       Set $\sigma^k := \sqrt{(d^k)^\top \widetilde{R}^k d^k}$, and go to Step 15.
11:     end if
12:     Update $\beta^k := \beta^k + c^{-1}((d^k)^\top \widetilde{R}^k d^k - (\sigma^k)^2)$.
13:     Project $\beta^k$ onto the interval $[0, 1]$, i.e. set $\beta^k := \max(\min(\beta^k, 1), 0)$.
14:   until the stopping criterion for the inner loop is fulfilled.
15:   Evaluate $\rho^k := [f_\gamma(u^k) - f_\gamma(u^k + d^k)] / [f_\gamma(u^k) - (f_\gamma(u^k) + (g^k)^\top d^k + (d^k)^\top \widetilde{H}^k d^k / 2)]$.
16:   if $\rho^k < \rho_1$ then
17:     Set $\sigma^{k+1} := \kappa_1 \sigma^k$.
18:   else if $\rho^k > \rho_2$ then
19:     Set $\sigma^{k+1} := \kappa_2 \sigma^k$.
20:   else
21:     $\sigma^{k+1} := \sigma^k$.
22:   end if
23:   Determine the step size $a^k$ along the search direction $d^k$ such that $u^{k+1} = u^k + a^k d^k$ satisfies the following Wolfe-Powell conditions:
          $f_\gamma(u^{k+1}) \leq f_\gamma(u^k) + \tau_1 a^k \nabla f_\gamma(u^k)^\top d^k,$          (2.8.15)
          $\nabla f_\gamma(u^{k+1})^\top d^k \geq \tau_2 \nabla f_\gamma(u^k)^\top d^k.$          (2.8.16)
24:   Generate the next iterate:
          $u^{k+1} := u^k + a^k d^k,$
          $p^{k+1} := \varphi(m^k)^{-1} \left( G u^k + (1 - \widetilde{p}^k \frac{\varphi'(m^k)(G u^k)}{m^k}) G d^k \right).$          (2.8.17)
```

```
25:     Initialize the $R$-regularization weight $\beta^{k+1} := \beta^k$ for the next iteration.
26:     Set $k := k + 1$.
27: until the stopping criterion for the outer loop is fulfilled.
```

Concerning the input parameters involved in the above algorithm, we note that these quantities are presented merely for the generality of the algorithm and do not require particular tuning for different test runs. Throughout our numerical experiments in section 2.8.3, we shall always fix the parameters as follows: $c = 1,\ \rho_1 = 0.25,\ \rho_2 = 0.75,\ \kappa_1 = 0.25,\ \kappa_2 = 2,\ \varepsilon = 10^{-4}\alpha,\ \epsilon_d = 10^{-8},\ \tau_1 = 0.1,\ \tau_2 = 0.9$.

We remark that Algorithm 2.8.8 is a hybrid approach combining the trust-region method and the line search technique. The Wolfe-Powell line search, along the search direction d^k obtained from the R-regularized Newton scheme, is responsible for the global convergence of the overall algorithm; see Theorem 2.8.9 in the following.

Theorem 2.8.9 (Global convergence)**.** *Let $\{(u^k, p^k)\}$ be the sequence generated by Algorithm 2.8.8. Then we have*

$$\lim_{k\to+\infty} \|\nabla f_\gamma(u^k)\| = 0. \tag{2.8.18}$$

Moreover, if in addition $\{u^k\}$ is uniformly bounded, then the sequence $\{(u^k, p^k)\}$ has an accumulation point (u^*, p^*) satisfying the Euler-Lagrange equation (2.8.5).

Proof. According to Theorem 2.2.14, we have $\sum_{k=0}^{\infty} \cos^2\theta^k \|g^k\|^2 < \infty$, where

$$\cos\theta^k := -\frac{(g^k)^\top d^k}{\|g^k\|\|d^k\|}.$$

Due to the descent property

$$\cos\theta^k \geq \epsilon_d > 0, \tag{2.8.19}$$

guaranteed by Theorem 2.8.5 and steps 6–8 in Algorithm 2.8.8, we have proved (2.8.18).

Moreover, it follows from the descent property (2.8.19) that

$$\epsilon_d \|g^k\|\|d^k\| \leq -(g^k)^\top d^k = (d^k)^\top(\widetilde{H}^k + \beta^k \widetilde{R}^k)d^k \leq \|g^k\|\|d^k\|.$$

Consider $d^k := s^k v^k$ such that $s^k \geq 0$ and $\|v^k\| = 1$ for all k, then we have

$$\epsilon_d \|g^k\| \leq s^k (v^k)^\top(\widetilde{H}^k + \beta^k \widetilde{R}^k)v^k \leq \|g^k\|.$$

It follows that

$$\lim_{k\to\infty} s^k (v^k)^\top(\widetilde{H}^k + \beta^k \widetilde{R}^k)v^k = 0. \tag{2.8.20}$$

By the uniform boundedness of $\{u^k\}$, $\{v^k\}$, $\{\widetilde{p}^k\}$, and $\{\beta^k\}$, there exist $u^*, v^* \in \mathbb{R}^{|\Omega_u|}$, $\widetilde{p}^* \in \mathbb{R}^{|\Omega_p|}$, and $\beta^* \in [0, 1]$ such that up to a subsequence $u^k \to u^*$, $v^k \to v^*$, $\widetilde{p}^k \to \widetilde{p}^*$, and $\beta^k \to \beta^*$ as $k \to \infty$. Owing to the continuity of the mappings $(u^k, \widetilde{p}^k) \mapsto \widetilde{H}^k = H(u^k, \widetilde{p}^k)$ and

$(u^k, \widetilde{p}^k) \mapsto \widetilde{R}^k = R(u^k, \widetilde{p}^k)$, we also have $\widetilde{H}^k \to \widetilde{H}^* := H(u^*, \widetilde{p}^*)$ and $\widetilde{R}^k \to \widetilde{R}^* := R(u^*, \widetilde{p}^*)$ as $k \to \infty$.

We claim that $\liminf_{k\to\infty} s^k = 0$. Assume the contrary that $\{s^k\}$ is uniformly bounded away from 0. Then because of (2.8.20) we have $(v^*)^\top(\widetilde{H}^* + \beta^*\widetilde{R}^*)v^* = 0$, or equivalently

$$(\widetilde{H}^* + \beta^*\widetilde{R}^*)v^* = 0,$$

due to the symmetry of the matrix. This leads to a contradiction as

$$\begin{aligned}
\epsilon_d &\leq -\frac{(g^k)^\top d^k}{\|g^k\|\|d^k\|} = \frac{(d^k)^\top(\widetilde{H}^k + \beta^k\widetilde{R}^k)d^k}{\|(\widetilde{H}^k + \beta^k\widetilde{R}^k)^{-1}d^k\|\|d^k\|} \\
&\leq \frac{((d^k)^\top(\widetilde{H}^k + \beta^k\widetilde{R}^k)d^k)\|(\widetilde{H}^k + \beta^k\widetilde{R}^k)d^k\|}{\|d^k\|^3} \\
&\leq \frac{\|(\widetilde{H}^k + \beta^k\widetilde{R}^k)d^k\|^2}{\|d^k\|^2} \\
&= \|(\widetilde{H}^k + \beta^k\widetilde{R}^k)v^k\|^2 \xrightarrow{k\to\infty} \|(\widetilde{H}^* + \beta^*\widetilde{R}^*)v^*\|^2 = 0.
\end{aligned}$$

We have used the Cauchy-Schwarz inequality in deriving the above inequalities. Thus, we have proved that $\liminf_{k\to\infty} \|d^k\| = 0$.

Upon extracting another subsequence of $\{d^k\}$ and using again the same notation for the indices, we have $\lim_{k\to\infty} d^k = 0$ and then

$$p^* := \lim_{k\to\infty} p^{k+1} = \varphi(\max(|Gu^*|, \gamma))^{-1}Gu^*,$$

according the update formula (2.8.17). Together with the already established fact that

$$0 = \lim_{k\to\infty} \nabla f_\gamma(u^k) = \nabla f_\gamma(u^*) = \nabla\Theta(u^*) + \alpha G^\top(\varphi(\max(|Gu^*|, \gamma))^{-1}Gu^*),$$

we conclude that (u^*, p^*) satisfies the Euler-Lagrange equation (2.8.5). □

In addition to the global convergence, the trust-region framework supplies a proper tuning of the R-regularization weight β^k, such that β^k will vanish asymptotically. Thus the algorithm converges locally at a superlinear rate to a local minimizer satisfying the second-order sufficient optimality condition (for semismooth problems); see Theorem 2.8.10 below. To sketch the proof, note that for sufficiently large k, $\widetilde{H}^k$ and $\widetilde{R}^k$ both become strictly positive definite. It follows that the alternating iterations on β^k and u^k, i.e. steps 4–14 of Algorithm 2.8.8, converge and therefore the Cauchy-point based model reduction criterion will be satisfied; see (2.4.23). Analogous to the classical trust-region method, the evaluation ratio ρ^k tends to 1 and the trust-region radius σ^k is uniformly bounded away from 0. As a result the weight β^k will vanish in the limit. Finally, the full step size $a^k = 1$ is admissible for all sufficiently k and the step d^k is asymptotically identical to a full semismooth Newton step. We refer to Theorem 2.4.10 for a complete proof of the local superlinear convergence.

Theorem 2.8.10 (Local convergence)**.** *Let the sequence* $\{(u^k, p^k)\}$ *generated by Algorithm 2.8.8 converge to some* (u^*, p^*) *satisfying the Euler-Lagrange equation (2.8.5). Assume that all generalized Hessians of* f_γ *at* u^* *are strictly positive definite. Then we have* $\lim_{k\to\infty} \beta^k = 0$ *and the sequence* $\{u^k\}$ *converges to* u^* *superlinearly, i.e.*

$$\|u^{k+1} - u^*\| = o(\|u^k - u^*\|), \quad \text{as } k \to \infty.$$

2.8.3 Selected applications

Here we present a numerical study of Algorithm 2.8.8. Throughout this subsection, the linear system in step 5 is handled by the conjugate gradient method with residual tolerance 0.01. Whenever, the matrix $\widetilde{H}^k + \beta^k \widetilde{R}^k$ is detected to be indefinite or (near-) singular, we immediately utilize step 7 in order to obtain a positive definite linear system. The trust-region subproblem (2.8.10)–(2.8.11) is solved only approximately. From our numerical experience, one (inner) iteration on the R-regularization weight β^k seems efficient for the overall algorithm. The regularization parameter α is manually chosen to properly balance the data fidelity and the sparsity-promoting prior. The Huber parameter γ is selected to be sufficiently small, depending on the particular application, in order to obtain a desirable sparse solution. We terminate the overall algorithm once the residual norm $\|\text{res}(u^k, p^k; \gamma)\|$ is reduced by a factor of 10^{-7} relative to its initial value.

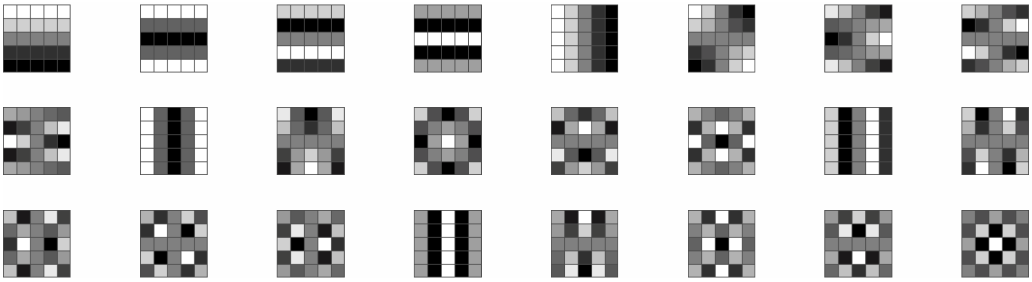
The remainder of this subsection will present selected applications of the general variational framework (2.8.1) in image processing, feature selection, and optimal control. All experiments in this subsection were performed under MATLAB R2011b on a 2.66 GHz Intel Core Laptop with 4 GB RAM. The reported CPU time is measured in seconds.

Image denoising via overcomplete dictionary

We first apply our method to an image denoising problem, where the following $\ell^{1/2}$-DCT5 model is considered

$$\min_{u\in\mathbb{R}^{|\Omega|}} \frac{1}{2}\|u - z\|^2 + \sum_{l=1}^{24} \sum_{(j_1,j_2)\in\Omega} \alpha_l \psi_\gamma(|(h_l * u)_{j_1,j_2}|). \tag{2.8.21}$$

Here, z is the observed image (see Figure 2.13(b)), which is generated by adding white Gaussian noise of standard deviation 25/255 to the “Cameraman” image (see Figure 2.13(a)). The filters $(h_l)_{l=1}^{24}$ are the two-dimensional 5th-order discrete cosine transform (DCT5) filters, and correspondingly $(\alpha_l)_{l=1}^{24}$ are the regularization parameters trained from a large database of image patches [KP13]; see Figure 2.12 for the illustrations of DCT5 filters and the values of trained regularization parameters. The symbol “$*$” denotes the conventional two-dimensional convolution. By considering the concave bridge prior with exponent 1/2 (or $\psi(t) = 2t^{1/2}$), we expect the restored image u to be sparse under the DCT5 transform. In this sense, the variational model (2.8.21) is an *analysis* approach [COS09].



5.599e-4	7.036e-4	4.913e-4	8.650e-4	9.291e-4	8.073e-4	9.853e-4	8.291e-4
1.981e-3	8.766e-4	6.595e-4	5.764e-4	7.636e-4	1.075e-3	9.150e-4	4.896e-4
6.361e-4	3.362e-4	1.180e-3	1.209e-3	1.392e-3	1.062e-3	2.121e-3	1.739e-3

Figure 2.12: DCT5 filters and regularization parameters.

We implement Algorithm 2.8.8 with the initial guess $u^0 = z$ and different choices of the Huber parameter, namely $\gamma = 0.03$, 0.02, 0.015, and 0.01. The quality of the restored image is measured by the peak signal-to-noise ratio (PSNR). The corresponding PSNR and CPU time with respect to different γ are reported in Table 2.7. We note the tradeoff in γ-selection that smaller γ typically yields the higher quality on the sparse solution, but costs more CPU time.

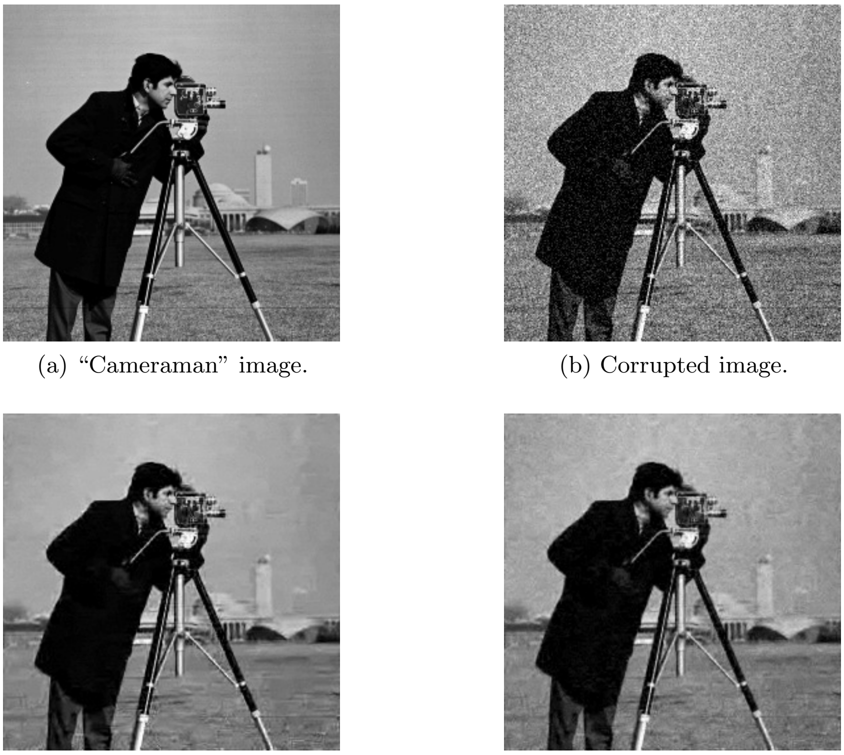
γ	0.03	0.02	0.015	0.01
PSNR	26.61	27.47	27.91	28.25
CPU	50.88	61.03	122.7	234.1

Table 2.7: Dependence on the Huber parameter.

We further compare the performance of the $\ell^{1/2}$-DCT5 model (with $u^0 = z$, $\gamma = 0.01$) and that of ℓ^1-DCT5 model in [KP13], for which the corresponding restored images are displayed in (c) and (d) of Figure 2.13, repectively. Table 2.8 reports the quantitative comparison of the two models with respect to the PSNR value, the number of Newton iterations, the total number of conjugate gradient iterations, and the CPU time. It is observed that the ℓ^1-DCT5 model poorly restores the homogeneous region in order to well preserve the textures in the image. The $\ell^{1/2}$-DCT5 model is more time-consuming due to solving a nonconvex problem, but yields considerable improvement on the restoration quality.

Model	PSNR	#Newton	#CG	CPU
ℓ^1-DCT5	27.46	11	104	80.37
$\ell^{1/2}$-DCT5	28.25	24	358	234.1

Table 2.8: Comparison of $\ell^{1/2}$- and ℓ^1-models.



(a) “Cameraman” image.

(b) Corrupted image.

(c) Restoration via $\ell^{1/2}$-DCT5.

(d) Restoration via ℓ^1-DCT5.

Figure 2.13: Image denoising via overcomplete dictionary.

Feature selection via sparse support vector machines

We consider an example of feature selection using a support vector machine (SVM) [WMC$^+$00], where we aim to identify 10 feature variables out of 200 candidate variables $(x_j)_{j=1}^{200}$. The identification is based on n training samples simulated as follows. For each sample, the outcome $y^i \in \{+1, -1\},\ i \in \{1, 2, ..., n\}$, is generated with equal probability. If x_j is a feature variable, then with probability 0.3 the random variable $x_j^i = y^i\mathcal{N}(3, 1)$ is drawn and with probability 0.7 we generate $x_j^i = \mathcal{N}(0, 1)$. If x_j is a noise variable, then $x_j^i = \mathcal{N}(0, 1)$ is independently generated.

The linear SVM uses the classifier $y = \mathrm{sgn}(b + \sum_{j=1}^{200} w_j x_j)$ to predict the outcome for a fresh input x. The unknowns $b \in \mathbb{R}$ and $w \in \mathbb{R}^{200}$ are determined by solving the following minimization problem

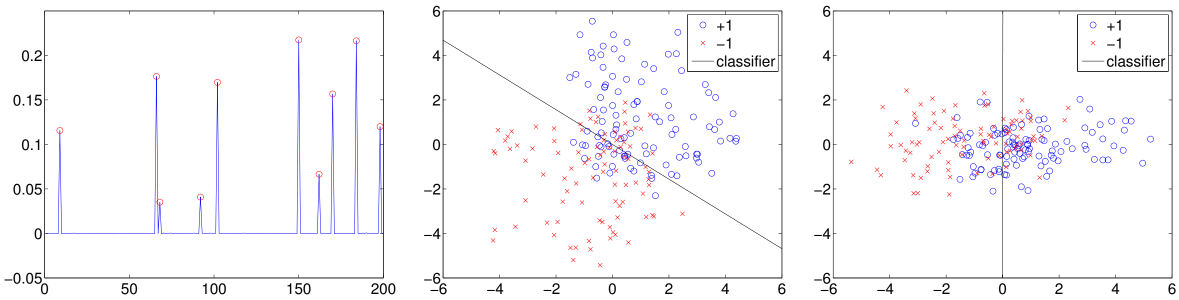
$$\min_{b\in\mathbb{R}, w\in\mathbb{R}^{200}} \ \alpha \sum_{j=1}^{200} \psi_\gamma(|w_j|) + \frac{1}{n}\sum_{i=1}^{n} L_{\epsilon_{hl}}\left(y^i(b + \sum_{j=1}^{200} w_j x_j^i)\right), \tag{2.8.22}$$

where $L_{\epsilon_{hl}}(\cdot)$ is a smoothed hinge loss [Cha07a] defined by

$$L_{\epsilon_{hl}}(s) = \begin{cases} \max(1 - s, 0), & \text{if } |s - 1| \geq \epsilon_{hl}, \\ (1 + \epsilon_{hl} - s)^2/(4\epsilon_{hl}), & \text{if } |s - 1| < \epsilon_{hl}, \end{cases}$$

with the smoothing parameter $\epsilon_{hl} = 0.01$. In this experiment, we choose $\alpha = 0.1$, $\psi(t) = \log(1 + 2t)$, and $\gamma = 0.001$.

The computational results for a trial run with $n = 200$ training samples are displayed in Figure 2.14. We plot the importance weight w in (a), where in particular the weights for the 10 presumed feature variables are marked by red circles. From this figure, it is observed that the variational model (2.8.22) has correctly identified the feature variables among all candidate variables. In (b) and (c), we illustrate the computed classifier with respect to the training data projected onto two particular candidate variables, i.e. $y = \text{sgn}(b + w_{j_1}x_{j_1} + w_{j_2}x_{j_2})$. More specifically, in (b) x_{j_1} and x_{j_2} are two distinct feature variables, and in (c) one is a feature variable and the other is a noise variable. In both figures, the coordinates of the circles indicate the random-variable values of those simulated samples with outcomes +1, and the coordinates of the crosses indicate the random-variable values of those simulated samples with outcomes −1.



(a) Importance weight w.

(b) Classification: feature-feature.

(c) Classification: feature-noise.

Figure 2.14: Feature selection via sparse support vector machine.

Sparse optimal control

Finally, we demonstrate an application in sparse optimal control, which shows considerable promises in actuator placement problems; see, e.g., [Sta09, CK12]. Consider the following stationary control problem:

$$\min \quad J(y,u) = \frac{1}{2}\int_\Omega |y - z|^2 dx + \frac{\mu}{2}\int_\Omega |\nabla u|^2 dx + \alpha \int_\Omega \psi(|u|)dx \tag{2.8.23}$$

$$\text{over } (y,u) \in H_0^1(\Omega) \times U_{ad}, \tag{2.8.24}$$

$$\text{s.t.} \quad \int_\Omega \nabla y \cdot \nabla v dx = \int_\Omega uv dx \quad \forall v \in H_0^1(\Omega). \tag{2.8.25}$$

Here Ω is a bounded Lipschitz domain, $\alpha > 0$, $0 < \mu \ll \alpha$ are some given parameters, a desired state is given by $z \in H_0^1(\Omega)$, and U_{ad} is some weakly closed subset in $H_0^1(\Omega)$. A (continuous) concave prior $\psi(\cdot)$ is applied in order to promote the sparsity of the optimal control in the spatial domain.

In general, it is a difficult task to establish the existence of solutions for a nonconvex minimization problem in function space due to the lack of weak (or weak*) lower semicontinuity for the objective; see, e.g., [AK02] and section 2.7 of the present chapter. However, in this special case with the H^1-regularization (the μ-term), we are able to show the existence of solution in the following theorem.

Theorem 2.8.11. *The stationary control problem (2.8.23)–(2.8.25) admits a solution.*

Proof. By the Lax-Milgram Lemma, the solution mapping $u \mapsto y = (-\Delta)^{-1}u$ for (2.8.25) is linear and continuous. Thus we only need to consider the reduced problem:

$$\min_{u \in U_{ad}} \widehat{J}(u) = \frac{1}{2}\int_\Omega |(-\Delta)^{-1}u - z|^2 dx + \frac{\mu}{2}\int_\Omega |\nabla u|^2 dx + \alpha \int_\Omega \psi(|u|)dx.$$

Since $\widehat{J}(\cdot)$ is bounded from below and coercive in $H_0^1(\Omega)$, any infimizing sequence $\{u^k\}$ is uniformly bounded in $H_0^1(\Omega)$. By the reflexivity of the space $H^1(\Omega)$ and the weak closedness of the admissible set U_{ad}, there exists a subsequence of $\{u^k\}$, also denoted by $\{u^k\}$, such that $u^k \rightharpoonup u^*$ in $H_0^1(\Omega)$ as $k \to \infty$ for some $u^* \in U_{ad}$.

As the functional $u \in H_0^1(\Omega) \mapsto \frac{1}{2}\int_\Omega |(-\Delta)^{-1}u - z|^2 dx + \frac{\mu}{2}\int_\Omega |\nabla u|^2 dx$ is convex and strongly continuous, it is weakly lower semicontinuous, and thus we have

$$\begin{aligned} &\frac{1}{2}\int_\Omega |(-\Delta)^{-1}u^* - z|^2 dx + \frac{\mu}{2}\int_\Omega |\nabla u^*|^2 dx \\ \le &\liminf_{k\to\infty} \frac{1}{2}\int_\Omega |(-\Delta)^{-1}u^k - z|^2 dx + \frac{\mu}{2}\int_\Omega |\nabla u^k|^2 dx. \end{aligned}$$

On the other hand, the compact embedding of $H_0^1(\Omega)$ into $L^2(\Omega)$ (see, e.g., Theorem 5.3.3 in [ABM06]) implies the strong convergence of $\{u^k\}$ to u^* in $L^2(\Omega)$, and thus we have, up to another subsequence, $u^k(x) \to u^*(x)$ a.e. in Ω as $k \to \infty$. By Fatou's lemma and the continuity of the scalar function $\psi(|\cdot|)$, we have

$$\int_\Omega \psi(|u^*|)dx \le \liminf_{k\to\infty} \int_\Omega \psi(|u^k|)dx.$$

Altogether, we have $\widehat{J}(u^*) \le \liminf_{k\to\infty} \widehat{J}(u^k)$, indicating that u^* is an optimal solution to the underlying problem. □

We remark that without the H^1-regularization term the above proof would no longer be valid due to the lack of coercivity of the reduced objective $\widehat{J}(\cdot)$. In addition, H^1-regularization enforces sufficient regularity on a weakly convergent (sub)sequence that finally yields the almost everywhere pointwise convergence of the infimizing (sub)sequence.

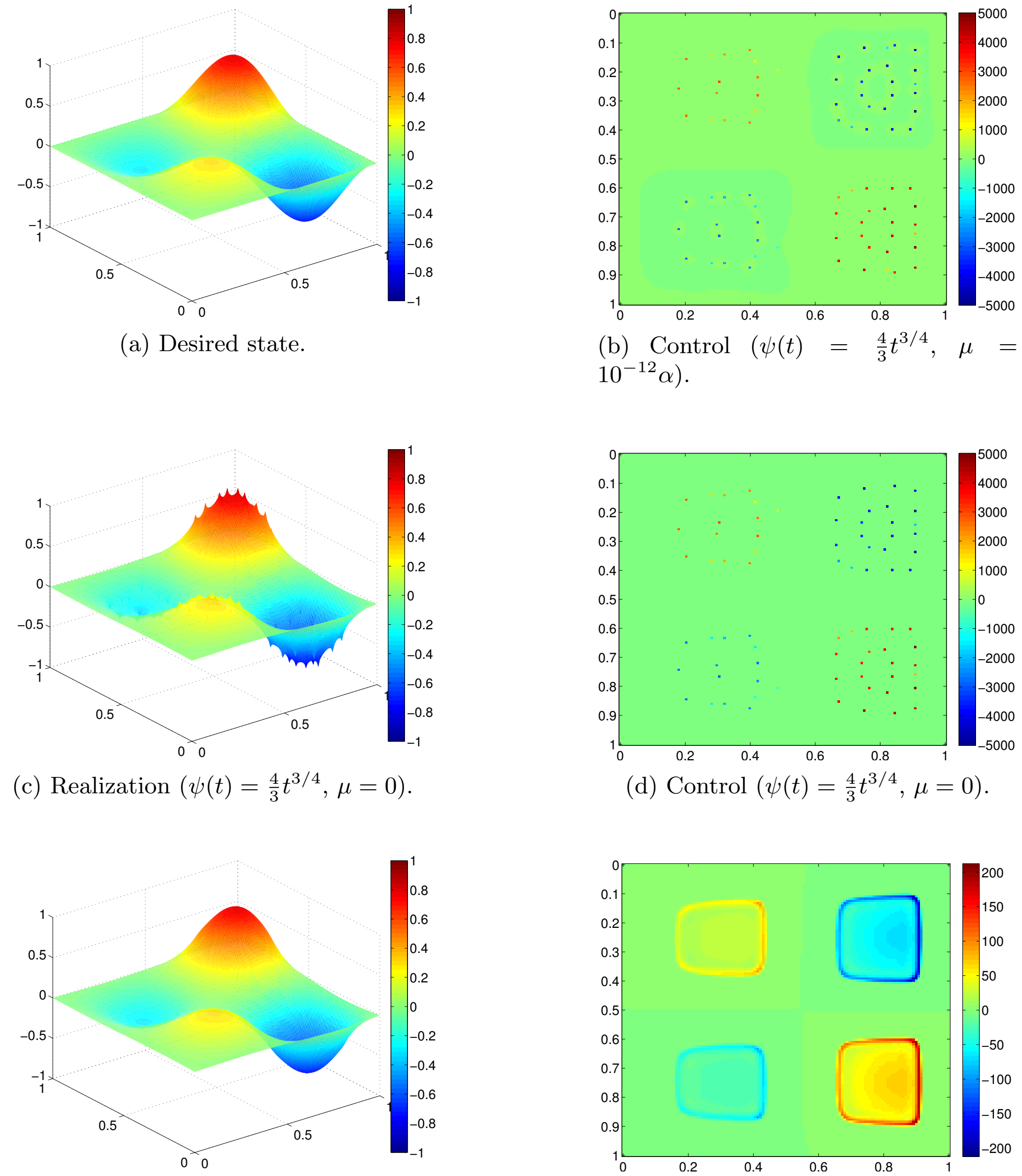
Now we turn our attention to the numerical solution for the following discretized control problem in reduced form:

$$\min_{u \in \mathbb{R}^{|\Omega|}} \sum_{(j_1,j_2)\in\Omega} \frac{1}{2}|(-\Delta)^{-1}u - z|^2 + \alpha\psi_\gamma(|u|) + \frac{\mu}{2}|\nabla u|^2.$$

Here $\Omega = \{0, 1, 2, ..., 2^N\}^2$, where $N \in \mathbb{N}$, denotes the 2D index set for the discretized square domain $(0, 1)^2$ with a uniform mesh size $h = 2^{-N}$. The Laplacian Δ with homogenous Dirichlet boundary conditions is discretized by the standard 5-point stencil. The desired state $z \in \mathbb{R}^{|\Omega|}$ is defined by

$$z_{j_1,j_2} = \sin(2\pi h j_1)\sin(2\pi h j_2)\exp(2hj_1)/6,$$

for all $(j_1, j_2) \in \Omega$; see Figure 2.15(a). Note that we have taken the admissible set to be universal, i.e. $U_{ad} = \mathbb{R}^{|\Omega|}$. In the following experiments, we shall fix $N = 7$, $\alpha = 10^{-4}$, $\gamma = 0.1$, and $u^0 = -\Delta z$. The associated numerical results are displayed in Figure 2.15.



(a) Desired state.

(b) Control ($\psi(t) = \frac{4}{3}t^{3/4}$, $\mu = 10^{-12}\alpha$).

(c) Realization ($\psi(t) = \frac{4}{3}t^{3/4}$, $\mu = 0$).

(d) Control ($\psi(t) = \frac{4}{3}t^{3/4}$, $\mu = 0$).

(e) Realization ($\psi(t) = t$, $\mu = 0$).

(f) Control ($\psi(t) = t$, $\mu = 0$).

Figure 2.15: Sparse optimal control.

As shown in (b), we compute the optimal control with the prior $\psi(t) = \frac{4}{3}t^{3/4}$ and $\mu = 10^{-12}\alpha$. In fact, in the discrete setting with fixed mesh size, this result is almost identical to the optimal control with $\mu = 0$ displayed in (d). Note that the optimal control is highly sparse in the spatial domain with sparsity rate $|\{(j_1, j_2) \in \Omega : |u_{j_1,j_2}| \geq \gamma\}|/|\Omega|$ equal to 0.47%. The corresponding realized state $(-\Delta)^{-1}u$ is given in (c), and the mean tracking error $\|(-\Delta)^{-1}u - z\|/|\Omega|$ is equal to 9.5322e-05. For comparison, we also compute the optimal control obtained from the (convex) prior $\psi(t) = t$ (together with $\mu = 0$), for which the realized state and the control are shown in (e) and (f), respectively. The corresponding sparsity rate of the control in (f) is 31.48% and the mean tracking error is 1.0041e-04. The comparison tells that the optimal control via the concave prior can produce a better realization of the desired state even with much higher spatial sparsity. Nevertheless, we remark that the magnitudes of the nontrivial entries (whose magnitudes are larger than γ) of the control in (d) are typically much larger than those in (f), which indicates that a higher physical capability of the control devices is typically required in order to compensate a reduction on the number of the control devices.

Chapter 3

Blind deconvolution via bilevel optimization

In this chapter, we continue our investigation on nonconvex and nonsmooth minimization approaches to sparse imaging, but in a somewhat different context from the previous chapter. Concretely, we shall investigate a bilevel optimization model for blind deconvolution, where the point spread function, which parameterizes the blurring operator, arises as a second unknown in addition to the underlying image.

3.1 Introduction

Image blur is widely encountered in many application areas; see, e.g., [CE07] and the references therein. In astronomy, images taken from a telescope appear blurry as light travels through a turbulent medium such as the atmosphere. The out-of-focus blur in microscopic images commonly occurs due to misplacement of the focal planes. Tomographic techniques in medical imaging, such as single-photon emission computed tomography (SPECT), are possibly prone to resolution limits of imaging devices or physical motion of patients, which both lead to blurring artifacts in final reconstructed images. In practice, the blurring operator, which can be modeled as the convolution with some *point spread function* (PSF) provided that the blurring is shift-invariant, is often not available beforehand and needs to be identified together with the underlying source image. Such a problem, typically known as *blind deconvolution* [KH96a, KH96b], represents an ill-posed inverse problem in image processing, more challenging than non-blind deconvolution owing to the coupling of the PSF and the image.

There exists a diverse literature on blind deconvolution, which roughly divides into two categories: direct methods and iterative methods. The direct methods, such as the APEX method by Carasso [Car01, Car02, Car06, Car09], typically assume a specific parametric structure on either the blurring kernel itself or its characteristic function, and are provably effective for specific applications. Among the iterative methods, some use simple fixed-point type iterations, e.g. the Richardson-Lucy method [FBP95], but their convergence properties and robustness against noise are difficult to analyze. Others proceed by formulating a proper variational model involving regularization terms on the image and/or the PSF. In [YK96] H^1-regularizations are imposed on both the image and the PSF, and in [CW98, HMO05] total-variation regularizations on the image and the PSF are utilized and yield better results than H^1-regularizations for certain PSFs. We also mention that nonconvex image priors are considered for blind deconvolution in the work [AA10], which are favorable for certain sparse images [CY08, HW13, HW14b]. The convergence analysis of an alternating minimization scheme for such double-regularization based variational approaches in appropriately chosen function spaces is carried out in [BS01, Jus06]. An exception of variational approaches to blind deconvolution is [JR06], where the optimality condition is "diagonalized" by Fourier transform and thus can be solved by some non-iterative root-finding algorithm. Although we shall focus ourselves only on spatially invariant PSFs in this work, we remark that blind deconvolution with spatially varying PSFs might be advantageous in certain applications such as telescopic imaging; see, e.g., [BJNP06].

Nevertheless, most existing variational approaches to blind deconvolution are "single-level", in the sense that both unknowns, i.e. the image and the PSF, appear in a single objective to be minimized. In this work, we are interested in a class of blind deconvolution problems where additional statistical information on the image (and possibly also on the PSF) is available. For instance, in microscopic imaging the blurring is nearly stationary and an artificial reference image can be inserted into the imaging device for obtaining a trial blurry observation of the reference image. In telescopic imaging, the target object, considered to be stationary, is photographed by multiple cameras within an instant, leading to highly correlated blurry observations. To exploit such additional image statistics, we propose a *bilevel optimization* framework. In essence, in the lower level the total-variation (TV) model (also known as the Rudin-Fatemi-Osher model [ROF92]) is imposed as the constraint that the underlying source image must comply with, as is typically done in non-blind deconvolution [AK02, CS05]. In the upper level, we minimize a suitable objective which incorporates the statistical information on the image and the PSF. Notably, bilevel optimization of similar structures has been recently applied to image processing for parameter/model learning tasks in [KP13, DlRS13].

Due to nonsmoothness of the objective in the (convex) TV-model, the sufficient and necessary optimality condition for the lower-level problem can be equivalently expressed as either a variational inequality, a nonsmooth equation, or a set-valued (or generalized) equation. This prevents us from applying the classical Karush-Kuhn-Tucker theory to derive a necessary optimality condition (or stationarity condition) for the overall bilevel optimization, and thus distinguishes our bilevel optimization problem from classical constrained optimization. Such difficulty is also typical in *mathematical programming with equilibrium constraints* (MPEC); see the monographs [LPR96, OKZ98] for comprehensive introductions on the subject. In this chapter, we tackle the total-variation based bilevel optimization problem in the fashion of MPEC. For the lower-level problem, we justify the so-called *strong regularity condition* by Robinson [Rob80] and then establish the B(ouligand)-differentiability of the solution mapping. Based on this, we derive the M(ordukhovich)-stationarity condition for the bilevel optimization problem. Yet, the C(larke)-stationarity, slightly weaker than the M-stationarity, is pursued numerically by a hybrid projected gradient method and its convergence is analyzed in detail. In the numerical experiments, we implement a simplified version of the hybrid projected gradient method and demonstrate some promising applications on point spread function calibration and multiframe blind deconvolution.

The rest of the chapter is organized as follows. Section 3.2 provides preliminaries on some classical theories concerning set-valued equations and MPECs. Then we formulate the bilevel optimization model in section 3.3. In section 3.4, the lower-level solution mapping is studied in detail with respect to its existence, continuity, and differentiability. Different notions of stationarity conditions are introduced in section 3.5, where their relations are also discussed. Section 3.6 develops and analyzes a hybrid projected gradient method for pursuing a C-stationary point of the bilevel problem. Numerical experiments based on a simplified project gradient method are presented in section 3.7.

3.2 Preliminaries on mathematical programs with equilibrium constraints (MPECs)

Consider the following MPEC in a general setting:

$$\begin{array}{rl} \text{minimize (min)} & J(u,h) \\ \text{subject to (s.t.)} & 0 \in F(u,h) + G(u), \\ & u \in \mathbb{R}^n,\ h \in Q_h. \end{array} \tag{3.2.1}$$

Here $Q_h \subset \mathbb{R}^m$ is a non-empty, closed admissible set for h, $J : \mathbb{R}^n \times Q_h \to \mathbb{R}$ is locally Lipschitz, $F : \mathbb{R}^n \times Q_h \to \mathbb{R}^n$ is continuously differentiable, and $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a set-valued mapping with a closed graph. Without loss of generality, we assume that there exists at least one feasible point (u, h) in (3.2.1) and that the MPEC problem (3.2.1) admits a global solution. Very often these conditions can be justified by standard arguments once further structural information on (3.2.1) is supplied. The goal of this preliminary section is to derive a stationarity condition for (3.2.1) based on Mordukhovich's generalized differential calculus.

3.2.1 Lower-level problem: strong regularity condition and an implicit map

We first focus on the set-valued equation

$$0 \in F(u,h) + G(u), \quad u \in \mathbb{R}^n,\ h \in Q_h. \tag{3.2.2}$$

which appears as a constraint in the MPEC (3.2.1). The solution mapping of (3.2.2) is denoted by $S : Q_h \rightrightarrows \mathbb{R}^n$ such that (3.2.2) holds for any $h \in Q_h$, $u \in S(h)$. We remark that the formulation in (3.2.2) covers a rich class of equilibrium problems including constrained minimizations, nonlinear complementarity problems, and variational inequalities (of the first kind); see [OKZ98] for a more comprehensive introduction.

In the following, we introduce the notation of strong regularity condition originally proposed by [Rob80], which leads to a (generalized) implicit function theorem for the set-valued equation.

Definition 3.2.1 (Strong regularity condition)**.** *Let $(u^0, h^0) \in \mathbb{R}^n \times Q_h$ be a reference solution for (3.2.2), i.e. $u^0 \in S(h^0)$, and $\Sigma : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be defined by*

$$\Sigma(\xi) := \left\{u \in \mathbb{R}^n : \xi \in F(u^0, h^0) + D_u F(u^0, h^0)(u - u^0) + G(u)\right\}. \tag{3.2.3}$$

Assume that there exist neighborhoods U_ξ of $0 \in \mathbb{R}^n$ and U_u of u^0 such that the mapping $\xi \mapsto \Sigma(\xi) \cap U_u$ is single-valued and Lipschitz on U_ξ. Then we say the strong regularity condition holds for (3.2.2) at (u^0, h^0).

Theorem 3.2.2 (Generalized implicit function theorem)**.** *Assume that the strong regularity condition in Definition 3.2.1 holds for (3.2.2) at a reference solution $(u^0, h^0) \in \operatorname{gph} S$. Then there exist neighborhoods V_h of h^0 and V_u of u^0 such that the mapping $h \in V_h \cap Q_h \mapsto S(h) \cap V_u$ is single-valued and Lipschitz.*

Proof. See Theorem 2.1 and Corollary 2.2 in [Rob80]. □

We refer to Chapter 5 in [OKZ98] for an exhibition of sufficient conditions that yield strong regularity condition for (3.2.2) when F and G admit particular structures.

3.2.2 Elements in Mordukhovich's generalized differential calculus

Here we collect important notations from Mordukhovich's generalized differential calculus and a few auxiliary results used later in section 3.2.3. The materials presented here are based on [Mor94, Out00]. For a more comprehensive introduction, we refer to the monographs [RW98, Mor06].

Definition 3.2.3 (Tangent and normal cones)**.** *The tangent (or contingent) cone of a subset Ξ in $\mathbb{R}^p$ at $a \in \operatorname{cl}\Xi$, denoted by $T_\Xi(a)$, is defined by*

$$T_\Xi(a) = \left\{v \in \mathbb{R}^p : t^k \to 0^+,\ v^k \to v,\ a + t^k v^k \in \Xi\ \forall k\right\}. \tag{3.2.4}$$

The (regular) normal cone of Ξ at $a \in \operatorname{cl}\Xi$, denoted by $N_\Xi(a)$, is defined as the (negative) polar cone of $T_\Xi(a)$, i.e.

$$N_\Xi(a) = \{w \in \mathbb{R}^p : \langle w, v\rangle \le 0\ \forall v \in T_\Xi(a)\}.$$

Definition 3.2.4 (Mordukhovich normal cone)**.** *The Mordukhovich (or limiting) normal cone of a subset Ξ in $\mathbb{R}^p$ at $a \in \operatorname{cl}\Xi$, denoted by $N_\Xi^{(\mathrm{M})}(a)$, is defined by*

$$N_\Xi^{(\mathrm{M})}(a) = \{w \in \mathbb{R}^p : w^k \to w,\ a^k \to a,\ w^k \in N_\Xi(a^k)\ \forall k\}.$$

Note that if Ξ is convex, we have $N_\Xi(\cdot) = N^{(\mathrm{M})}_\Xi(\cdot)$; see [Mor94]. In the following, for a set-valued mapping $\Psi : \mathbb{R}^p \rightrightarrows \mathbb{R}^q$, we denote its graph by $\operatorname{gph}\Psi := \{(a, b) \in \mathbb{R}^p \times \mathbb{R}^q : b \in \Psi(a)\}$. For an extended real-valued function $\psi : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, we denote its epigraph by $\operatorname{epi}\psi := \{(a, b) \in \mathbb{R}^p \times \mathbb{R} : b \geq \psi(a)\}$. Without further specification, Ψ is a generic set-valued mapping with a closed graph in our discussion.

Definition 3.2.5 (Coderivative)**.** *The coderivative of a set-valued mapping $\Psi : \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ at $(a, b) \in \operatorname{gph}\Psi$ is a set-valued mapping $D^*\Psi(a, b) : \mathbb{R}^q \rightrightarrows \mathbb{R}^p$ defined by*

$$D^*\Psi(a, b)[\delta b] := \left\{\delta a \in \mathbb{R}^p : (\delta a, -\delta b) \in N^{(\mathrm{M})}_{\operatorname{gph}\Psi}(a, b)\right\}.$$

Definition 3.2.6 (Mordukhovich subdifferential)**.** *The Mordukhovich subdifferential, denoted by $\partial^*\psi(a)$, of an extended real-valued function $\psi : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ at $a \in \mathbb{R}^p$ is defined by*

$$\partial^*\psi(a) := \left\{\delta a \in \mathbb{R}^p : (\delta a, -1) \in N^{(\mathrm{M})}_{\operatorname{epi}\psi}(a, \psi(a))\right\}.$$

Lemma 3.2.7. *Assume that $\Psi_1 : \mathbb{R}^p \to \mathbb{R}^q$ is continuously differentiable and $\Psi_2 : \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ has a closed graph. Then for any $b \in \Psi_1(a) + \Psi_2(a)$ and $\delta b \in \mathbb{R}^q$, we have*

$$D^*(\Psi_1 + \Psi_2)(a, b)[\delta b] = D\Psi_1(a)^\top \delta b + D^*\Psi_2(a, b - \Psi_1(a))[\delta b].$$

Proof. See Corollary 4.4 in [Mor94]. □

Lemma 3.2.8. *Let Q_1, Q_2 be two subsets in $\mathbb{R}^p$ and $a \in \operatorname{cl} Q_1 \cap \operatorname{cl} Q_2$. If $N^{(\mathrm{M})}_{Q_1}(a) \cap \left(-N^{(\mathrm{M})}_{Q_2}(a)\right) = \{0\}$, then the following inclusion holds true:*

$$N^{(\mathrm{M})}_{Q_1 \cap Q_2}(a) \subset N^{(\mathrm{M})}_{Q_1}(a) + N^{(\mathrm{M})}_{Q_2}(a).$$

Proof. See Corollary 4.7 in [Mor94]. □

Lemma 3.2.9. *Consider a continuously differentiable function $H : \mathbb{R}^p \to \mathbb{R}^q$ and a nonempty, closed subset Q_H in $\mathbb{R}^q$. Define the set-valued mapping $\Phi : \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ by*

$$\Phi(a) := H(a) + Q_H \quad \forall a \in \mathbb{R}^p. \tag{3.2.5}$$

Then the coderivative of Φ at $(a, b) \in \operatorname{gph}\Phi$ is given by

$$D^*\Phi(a, b)[\delta b] = \begin{cases} DH(a)^\top \delta b & \text{if } \delta b \in -N^{(\mathrm{M})}_{Q_H}(b - H(a)), \\ \emptyset & \text{otherwise.} \end{cases} \tag{3.2.6}$$

Proof. See Lemma 2.3 in [Out00]. □

Definition 3.2.10 (Pseudo-Lipschitz continuity)**.** *A set-valued mapping $\Psi : \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ is pseudo-Lipschitz continuous at $(a^0, b^0) \in \operatorname{gph}\Psi$ with modulus $L_\Psi \geq 0$ if there exist neighborhoods U_a of a^0 and U_b of b^0 such that the following inclusion holds for all $a \in U_a$:*

$$\Psi(a) \cap U_b \subset \Psi(a^0) + L_\Psi \|a - a^0\| \mathbb{B},$$

where $\mathbb{B}$ is the closed unit ball in $\mathbb{R}^q$.

Lemma 3.2.11. *Consider a set-valued mapping $\Psi : \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ and $(a, b) \in \operatorname{gph} \Psi$. Then Ψ is pseudo-Lipschitz continuous at (a, b) if and only if*

$$D^* \Psi(a, b)[0] = \{0\}.$$

Proof. See Proposition 2.8 in [Mor94]. □

3.2.3 Mordukhovich-type stationarity condition for MPECs

Now we apply the results from section 3.2.2 to derive a stationarity condition for the MPEC (3.2.1). Let us first consider a single-level mathematical program as follows:

$$\min \ \psi(a) \quad \text{s.t. } a \in \Xi, \tag{3.2.7}$$

where we assume $\psi : \mathbb{R}^p \to \mathbb{R}$ is locally Lipschitz and Ξ is a nonempty, closed subset in $\mathbb{R}^p$.

Lemma 3.2.12. *Any local solution a^* for (3.2.7) satisfies*

$$0 \in \partial^* \psi(a^*) + N_{\Xi}^{(\mathrm{M})}(a^*).$$

Proof. See Proposition 2.1 in [Out00]. □

Now we assume that the constraint adopts the following representation

$$\Xi = \{a \in Q_a : 0 \in H(a) + Q_H\}, \tag{3.2.8}$$

with a continuously differentiable function $H : \mathbb{R}^p \to \mathbb{R}^q$ and nonempty, closed subsets $Q_a \subset \mathbb{R}^p$, $Q_H \subset \mathbb{R}^q$. The following lemma relates (3.2.7) to a penalized program based on a pseudo-Lipschitz condition.

Lemma 3.2.13. *Assume that a^* is a local solution for (3.2.7) with Ξ given by (3.2.8) and ψ is Lipschitz near a^* with modulus L_ψ. Define the set-valued mapping $\Phi : \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ by*

$$\Phi(a) := H(a) + Q_H \quad \forall a \in \mathbb{R}^p, \tag{3.2.9}$$

and assume that $\Phi^{-1} \cap Q_a$ is pseudo-Lipschitz continuous at $(0, a^*)$ with modulus L_Φ. Then there exist neighborhoods U_b of $0 \in \mathbb{R}^q$ and U_a of a^* such that $(a, b) = (a^*, 0)$ solves the penalized program:

$$\begin{array}{ll} \min & \psi(a) + \lambda \|b\| \\ \text{s.t.} & b \in \Phi(a) \cap U_b, \ a \in Q_a \cap U_a, \end{array} \tag{3.2.10}$$

provided that $\lambda \geq L_\psi L_\Phi$.

Proof. See Lemma 2.2 in [Out00]. □

The follow theorem establishes a Mordukhovich-type stationarity condition for (3.2.7)–(3.2.8), which is taken from Theorem 2.4 in [Out00]. For completeness, we present a proof for this fundamental result.

Theorem 3.2.14. *Under the same assumption as in Lemma 3.2.13, there exists a multiplier $v^* \in N^{(\mathrm{M})}_{Q_H}(-H(a^*))$ such that*

$$0 \in \partial^* \psi(a^*) - DH(a^*)^\top v^* + N^{(\mathrm{M})}_{Q_a}(a^*). \tag{3.2.11}$$

Proof. Consider the penalized program (3.2.10) in Lemma 3.2.13 for sufficiently large λ. Then according to Lemma 3.2.12, the following stationarity condition holds true:

$$0 \in \partial^* \psi(a^*) \times (\lambda \mathbb{B}) + N^{(\mathrm{M})}_{\mathrm{gph}\, \Phi \,\cap\, (U_a \times U_b) \,\cap\, (Q_a \times \mathbb{R}^q)}(a^*, 0). \tag{3.2.12}$$

By Definition 3.2.5, we have

$$N^{(\mathrm{M})}_{\mathrm{gph}\, \Phi \,\cap\, (U_a \times U_b)}(a^*, 0) = \{(\delta a, \delta b) \in \mathbb{R}^p \times \mathbb{R}^q : \delta a \in D^* \Phi(a^*, 0)[-\delta b]\}.$$

Due to Lemma 3.2.9, this further implies

$$N^{(\mathrm{M})}_{\mathrm{gph}\, \Phi \,\cap\, (U_a \times U_b)}(a^*, 0) = \Big\{(\delta a, \delta b) \in \mathbb{R}^p \times \mathbb{R}^q : \delta a = -DH(a^*)^\top \delta b,\ \delta b \in N^{(\mathrm{M})}_{Q_H}(-H(a^*))\Big\}.$$

On the other hand, due to separability we have

$$N^{(\mathrm{M})}_{Q_a \times \mathbb{R}^q}(a^*, 0) = N^{(\mathrm{M})}_{Q_a}(a^*) \times \{0\}.$$

Thus, the assumption in Lemma 3.2.8 is fulfilled, i.e.

$$N^{(\mathrm{M})}_{\mathrm{gph}\, \Phi \,\cap\, (U_a \times U_b)}(a^*, 0) \cap \Big(-N^{(\mathrm{M})}_{Q_a \times \mathbb{R}^q}(a^*, 0)\Big) = \{0\}.$$

Invoking Lemma 3.2.8 on (3.2.12), we conclude that (3.2.11) holds for some $v^* \in N^{(\mathrm{M})}_{Q_H}(-H(a^*))$. □

To put the MPEC (3.2.1) into the perspective of Theorem 3.2.14, one may consider $a := (u, h) \in \mathbb{R}^n \times \mathbb{R}^m$, $H(a) := (-u, F(u, h))$, $Q_a := \mathbb{R}^n \times Q_h$, $Q_H := \mathrm{gph}\, G$. Accordingly, Φ in (3.2.9) takes the form

$$\Phi(u, h) = \begin{bmatrix} -u \\ F(u, h) \end{bmatrix} + \mathrm{gph}\, G. \tag{3.2.13}$$

Lemma 3.2.15. *Let (u^*, h^*) be a local solution for (3.2.1). Further assume the following constraint qualification (CQ):*

$$\left.\begin{array}{l} 0 \in \begin{bmatrix} I & -D_u F(u^*, h^*)^\top \\ 0 & -D_h F(u^*, h^*)^\top \end{bmatrix} \begin{bmatrix} \zeta \\ \eta \end{bmatrix} + \{0\} \times N^{(\mathrm{M})}_{Q_h}(h^*) \\ (\zeta, \eta) \in N^{(\mathrm{M})}_{\mathrm{gph}\, G}(u^*, -F(u^*, h^*)) \end{array}\right\} \Rightarrow \left\{ \begin{array}{l} \zeta = 0 \\ \eta = 0 \end{array}\right. . \tag{3.2.14}$$

Then there exist multipliers $\zeta^*, \eta^* \in \mathbb{R}^n$ such that the following stationarity condition is satisfied:

$$\begin{cases} 0 \in \partial^* J(u^*, h^*) + \begin{bmatrix} I & -D_u F(u^*, h^*)^\top \\ 0 & -D_h F(u^*, h^*)^\top \end{bmatrix} \begin{bmatrix} \zeta^* \\ \eta^* \end{bmatrix} + \{0\} \times N^{(\mathrm{M})}_{Q_h}(h^*), \\ (\zeta^*, \eta^*) \in N^{(\mathrm{M})}_{\mathrm{gph}\, G}(u^*, -F(u^*, h^*)). \end{cases} \tag{3.2.15}$$

Proof. See Theorem 3.1 in [Out00]. □

In fact, the CQ in (3.2.14) corresponds to the well-known Mangasarian-Fromowitz CQ on the constraint set $\{(u,h) \in \mathbb{R}^n \times Q_h : 0 \in \Phi(u,h),\ \Phi \text{ given by (3.2.13)}\}$ in the dual form. It is asserted in the following theorem that this CQ can be fulfilled by the strong regularity condition. The proof essentially follows Proposition 3.2 in [Out00].

Theorem 3.2.16. *Let (u^*, h^*) be a local solution for (3.2.1). Further assume that the strong regularity condition, see Definition 3.2.1, holds for the lower-level problem (3.2.2) at (u^*, h^*). Then the CQ in (3.2.14) is fulfilled and, therefore, the stationarity condition (3.2.15) must hold.*

Proof. Due to the strong regularity, the set-valued mapping Σ in (3.2.3) is pseudo-Lipschitz continuous at $(0, u^*)$. It follows from Lemma 3.2.11 that

$$D^*\Sigma(0,u^*)[0] = \{0\}. \tag{3.2.16}$$

In view of Definition 3.2.5, one can readily verify that

$$\delta u \in D^*\Sigma^{-1}(u^*,0)[\delta\xi] \;\Leftrightarrow\; -\delta\xi \in D^*\Sigma(0,u^*)[-\delta u],$$

and thus derive from (3.2.16) that

$$\operatorname{Ker} D^*\Sigma^{-1}(u^*,0) = \{0\}. \tag{3.2.17}$$

Invoking the summation rule in Lemma 3.2.7 on Σ^{-1}, we have

$$D^*\Sigma^{-1}(u^*,0)[-\eta] = D_uF(u^*,h^*)^\top(-\eta) + D^*G(u^*, -F(u^*,h^*))[-\eta].$$

This, together with (3.2.17), yields the following implication:

$$\left.\begin{array}{l} -D_uF(u^*,h^*)^\top\eta + \zeta = 0 \\ (\zeta,\eta) \in N^{(\mathrm{M})}_{\operatorname{gph} G}(u^*, -F(u^*,h^*)) \end{array}\right\} \;\Rightarrow\; \eta = 0,$$

which ensures the satisfaction of the CQ in (3.2.14). □

3.3 A bilevel optimization model for blind deconvolution

Let $u_{(\text{true})} \in \mathbb{R}^{|\Omega_u|}$ be the underlying source image over some two-dimensional (2D) index domain Ω_u. Assume the following image formation model for a blurry observation $z \in \mathbb{R}^{|\Omega_u|}$:

$$z = K(h_{(\text{true})})u_{(\text{true})} + \text{noise}. \tag{3.3.1}$$

Here the noise is assumed to be white Gaussian noise. We denote by $\mathcal{L}(\mathbb{R}^{|\Omega_u|})$ the set of all continuous linear maps from $\mathbb{R}^{|\Omega_u|}$ to itself and assume that $K : h \in Q_h \mapsto K(h) \in \mathcal{L}(\mathbb{R}^{|\Omega_u|})$ is a given continuously differentiable mapping over a convex and compact domain Q_h in $\mathbb{R}^m$.

In our theoretical and algorithmic development each $K(h)$ is only required to be a continuous linear operator on $\mathbb{R}^{|\Omega_u|}$, while in our numerics we focus on the cases where $K(h)$ represents a 2D convolution with some point spread function h, denoted by $K(h)u := h * u$. Thus, our task is to restore both unknowns, $u_{(\text{true})}$ and $h_{(\text{true})}$, from the observation z.

Whenever h is given, restoration of u (as non-blind deconvolution) can be carried out by solving the following variational problem:

$$\text{minimize } \frac{\mu}{2}\|\nabla u\|^2 + \frac{1}{2}\|K(h)u - z\|^2 + \alpha\|\nabla u\|_1 \quad \text{over } \ u \in \mathbb{R}^{|\Omega_u|}, \tag{3.3.2}$$

for some manually chosen parameters $\alpha > 0$ and $0 \leq \mu \ll \alpha$. Here $\nabla : \mathbb{R}^{|\Omega_u|} \to \left(\mathbb{R}^{|\Omega_u|}\right)^2$ is the discrete gradient operator, and we shall denote the discrete Laplacian by $\Delta := -\nabla^\top\nabla$. Besides, $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^{|\Omega_u|}$ or $\left(\mathbb{R}^{|\Omega_u|}\right)^2$, and $\|\cdot\|_1$ is the ℓ^1-norm defined by $\|p\|_1 := \sum_{j\in\Omega_u}|p_j|$ for $p \in \left(\mathbb{R}^{|\Omega_u|}\right)^2$ where each $|p_j|$ is the Euclidean norm of the vector $p_j \in \mathbb{R}^2$. We also denote by $\langle\cdot,\cdot\rangle$ the standard inner product in $\mathbb{R}^2$, $\mathbb{R}^{|\Omega_u|}$, or $\left(\mathbb{R}^{|\Omega_u|}\right)^2$.

The variational model (3.3.2) represents a discrete version of the Hilbert-space approach [IK99, HS06] to total variation (TV) image restoration:

$$\text{minimize } \int_\Omega \left(\frac{\mu}{2}|\nabla u|^2 + \frac{1}{2}|K(h)u - z|^2 + \alpha|\nabla u|\right) dx \quad \text{over } \ u \in H_0^1(\Omega).$$

Throughout this chapter, we shall assume for all feasible $h \in Q_h$ that

$$\text{Ker}\nabla \cap \text{Ker}K(h) = \{0\}, \tag{3.3.3}$$

or equivalently that $-\mu\Delta + K(h)^\top K(h)$ is positive definite. This assumption indicates that $K(h)$, for any $h \in Q_h$, does not annihilate constant vectors, as is indeed the case for the convolution with commonly encountered point spread functions. Provided that (3.3.3) holds true, the problem (3.3.2) always admits a unique global minimizer due to the strict convexity of the objective, for which the sufficient and necessary optimality condition is given by the following set-valued equation:

$$0 \in F(u,h) + G(u), \tag{3.3.4}$$

where $F : \mathbb{R}^{|\Omega_u|} \times Q_h \to \mathbb{R}^{|\Omega_u|}$ and $G : \mathbb{R}^{|\Omega_u|} \rightrightarrows \mathbb{R}^{|\Omega_u|}$ are respectively defined as

$$F(u,h) = (-\mu\Delta + K(h)^\top K(h))u - K(h)^\top z, \tag{3.3.5}$$

$$G(u) = \left\{\alpha\nabla^\top p : p \in (\mathbb{R}^{|\Omega_u|})^2, \ \begin{cases} p_j = \frac{(\nabla u)_j}{|(\nabla u)_j|} & \text{if } j \in \Omega_u, \ (\nabla u)_j \neq 0 \\ |p_j| \leq 1 & \text{if } j \in \Omega_u, \ (\nabla u)_j = 0 \end{cases}\right\}. \tag{3.3.6}$$

We remark that in the original work by Robinson [Rob80] the term generalized equations was used for set-valued equations.

In this work, we propose a bilevel optimization approach to blind deconvolution. In an abstract setting, the corresponding model reads

$$\begin{array}{rl} \min & J(u,h) \\ \text{s.t.} & 0 \in F(u,h) + G(u), \\ & u \in \mathbb{R}^{|\Omega_u|}, \ h \in Q_h. \end{array} \tag{3.3.7}$$

Here the total-variation model (3.3.2) represents the *lower-level problem* equivalently formulated as the first-order optimality condition (3.3.4), while in the *upper-level problem* we minimize a given objective $J : \mathbb{R}^{|\Omega_u|} \times Q_h \to \mathbb{R}$ known to be continuously differentiable and bounded from below. In this context, the set-valued equation (3.3.4) may be referred to as the *state equation* for the bilevel optimization (3.3.7), which implicitly induces a parameter-to-state mapping, i.e. $h \mapsto u$.

3.4 Solution mapping for lower-level problem: existence, continuity, and differentiability

In this section, we investigate the solution mapping associated with the lower-level problem in (3.3.7). To begin with, we establish the existence of such a solution mapping and its Lipchitz property by following Robinson's approach to set-valued equations [Rob80]. In this context, the notion of the strong regularity condition [Rob80] plays an important role. Essentially, the strong regularity condition for set-valued equations generalizes the invertibility condition in the classical implicit function theorem (for single-valued equations), and thus allows the application of Robinson's generalized implicit function theorem; see Theorem 3.2.2. In Theorem 3.4.1, we justify the strong regularity condition at any feasible point and its consequence turns out to be far-reaching. In what follows, we write $D_u F(u, h)$ for the (partial) differential of F with respect to u.

Theorem 3.4.1 (Strong regularity and implicit function)**.** *The strong regularity condition, see Definition 3.2.1, holds at any feasible solution (u^0, h^0) of (3.3.4), i.e. the mapping $w \in \mathbb{R}^{|\Omega_u|} \mapsto \{u \in \mathbb{R}^{|\Omega_u|} : w \in F(u^0, h^0) + D_u F(u^0, h^0)(u - u^0) + G(u)\}$ is (globally) single-valued and Lipschitz continuous. Consequently, there exists a locally Lipschitz continuous solution mapping $S : h \mapsto u$ such that $u = S(h)$ satisfies the set-valued equation (3.3.4) for all h.*

Proof. Due to Theorem 3.2.2, it suffices to show that the mapping $w \mapsto \{u \in \mathbb{R}^{|\Omega_u|} : w \in F(u^0, h^0) + D_u F(u^0, h^0)(u - u^0) + G(u)\}$ is globally single-valued and Lipschitz continuous.

First, note that $F(u^0, h^0) + D_u F(u^0, h^0)(u - u^0) = (-\mu\Delta + K(h^0)^\top K(h^0))u - K(h^0)^\top z$. Then the single-valuedness follows directly from the fact that the mapping

$$0 \in (-\mu\Delta + K(h^0)^\top K(h^0))u - K(h^0)^\top z - w + G(u)$$

is the sufficient and necessary condition for the (strictly) convex minimization

$$\min_u \frac{\mu}{2}\|\nabla u\|^2 + \frac{1}{2}\|K(h^0)u - z\|^2 - \langle w, u\rangle + \alpha\|\nabla u\|_1,$$

which admits a unique solution.

To prove the Lipschitz property, consider pairs (u^1, w^1) and (u^2, w^2) that satisfy

$$0 \in (-\mu\Delta + K(h^0)^\top K(h^0))u^1 - K(h^0)^\top z - w^1 + G(u^1),$$
$$0 \in (-\mu\Delta + K(h^0)^\top K(h^0))u^2 - K(h^0)^\top z - w^2 + G(u^2).$$

Then there exist subdifferentials $p^1 \in \partial \| \cdot \|_1(\nabla u^1)$ and $p^2 \in \partial \| \cdot \|_1(\nabla u^2)$ such that

$$\begin{aligned}
0 &= (-\mu\Delta + K(h^0)^\top K(h^0))u^1 - K(h^0)^\top z - w^1 + \alpha\nabla^\top p^1, \\
0 &= (-\mu\Delta + K(h^0)^\top K(h^0))u^2 - K(h^0)^\top z - w^2 + \alpha\nabla^\top p^2.
\end{aligned}$$

It follows from the property of subdifferentials in convex analysis, see e.g. Proposition 8.12 in [RW98], that

$$\begin{aligned}
\|\nabla u^2\|_1 &\geq \|\nabla u^1\|_1 + \langle p^1, \nabla u^2 - \nabla u^1 \rangle, \\
\|\nabla u^1\|_1 &\geq \|\nabla u^2\|_1 + \langle p^2, \nabla u^1 - \nabla u^2 \rangle,
\end{aligned}$$

which further implies that

$$\langle p^1 - p^2, \nabla u^1 - \nabla u^2 \rangle \geq 0.$$

Thus, we have

$$\begin{aligned}
0 &= \langle (-\mu\Delta + K(h^0)^\top K(h^0))(u^1 - u^2) - (w^1 - w^2) + \alpha\nabla^\top (p^1 - p^2), u^1 - u^2 \rangle \\
&\geq \langle (-\mu\Delta + K(h^0)^\top K(h^0))(u^1 - u^2), u^1 - u^2 \rangle - \langle w^1 - w^2, u^1 - u^2 \rangle,
\end{aligned}$$

and therefore the following Lipschitz property holds, i.e.

$$\|u^1 - u^2\| \leq \frac{1}{\lambda_{\min}(-\mu\Delta + K(h^0)^\top K(h^0))} \|w^1 - w^2\|,$$

where $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue of a matrix. This completes the proof. □

In view of Theorem 3.4.1, we may conveniently consider the reduced problem

$$\begin{array}{ll} \min & \widehat{J}(h) := J(u(h), h) \\ \text{s.t.} & h \in Q_h, \end{array} \tag{3.4.1}$$

which is equivalent to (3.3.7). It is immediately observed from (3.4.1) that there exists a global minimizer for (3.4.1) and thus also for (3.3.7).

Note that the state equation (3.3.4) can be expressed in terms of (u, h, p) as follows:

$$\left\{ \begin{array}{l} F(u, h) + \alpha\nabla^\top p = 0, \\ (u, \alpha\nabla^\top p) \in \operatorname{gph} G, \end{array} \right. \tag{3.4.2}$$

where p is included as an auxiliary variable lying in the set

$$Q_p := \left\{ p \in (\mathbb{R}^{|\Omega_u|})^2 : |p_j| \leq 1 \ \forall j \in \Omega_u \right\},$$

and $\operatorname{gph} G$ denotes the graph of the set-valued mapping G, i.e. $\operatorname{gph} G = \{(u, v) : u \in \mathbb{R}^{|\Omega_u|}, v \in G(u)\}$. We call the triplet (u, h, p) a *feasible point* for (3.3.7) if (u, h, p) satisfies (3.4.2).

In the following, we briefly introduce notions from variational geometry such as tangent/normal cones and graphical derivatives. The interested reader may find further details in Chapter 6 of the monograph [RW98]. Recall that the definitions of tangent and normal cones are given in Definition 3.2.3. In our context, the tangent and normal cones of $\operatorname{gph} G$ can be progressively calculated as:

$$
\begin{aligned}
T_{\operatorname{gph} G}(u, \alpha\nabla^\top p) = \Bigg\{ & (\delta u, \alpha \nabla^\top \delta p) : \\
& \left\{\begin{array}{ll} |(\nabla u)_j| \delta p_j = (\nabla \delta u)_j - \langle (\nabla \delta u)_j, p_j \rangle p_j & \text{if } (\nabla u)_j \neq 0, \\ (\nabla \delta u)_j = 0,\ \delta p_j \in \mathbb{R}^2 & \text{if } |p_j| < 1, \\ (\nabla \delta u)_j = 0,\ \langle \delta p_j, p_j \rangle \leq 0,\ \text{or} & \\ (\nabla \delta u)_j = c p_j\ (c \geq 0),\ \langle \delta p_j, p_j \rangle = 0 & \text{if } (\nabla u)_j = 0,\ |p_j| = 1. \end{array}\right. \Bigg\},
\end{aligned} \tag{3.4.3}
$$

$$
\begin{aligned}
N_{\operatorname{gph} G}(u, \alpha\nabla^\top p) = \Bigg\{ & (\alpha \nabla^\top w, -v) : \\
& \left\{\begin{array}{ll} w_j = \xi_j - \langle \xi_j, p_j \rangle p_j,\ (\nabla v)_j = |(\nabla u)_j| \xi_j\ (\xi_j \in \mathbb{R}^2) & \text{if } (\nabla u)_j \neq 0, \\ w_j \in \mathbb{R}^2,\ (\nabla v)_j = 0 & \text{if } |p_j| < 1, \\ \langle w_j, p_j \rangle \leq 0,\ (\nabla v)_j = c p_j\ (c \leq 0) & \text{if } (\nabla u)_j = 0,\ |p_j| = 1. \end{array}\right. \Bigg\}.
\end{aligned} \tag{3.4.4}
$$

The directional differentiability of the solution mapping S invokes the following notion.

Definition 3.4.2 (Graphical derivative)**.** *Let $S : V \rightrightarrows W$ be a set-valued mapping between two normed vector spaces V and W. The graphical derivative of S at $(v, w) \in \operatorname{gph} S$, denoted by $DS(v, w)$, is a set-valued mapping from V to W such that $\operatorname{gph} DS(v, w) = T_{\operatorname{gph} S}(v, w)$, i.e.*

$$\delta w \in DS(v, w)[\delta v] \quad \text{if and only if } (\delta v, \delta w) \in T_{\operatorname{gph} S}(v, w).$$

Notably, when S is single-valued and locally Lipchitz near $(v, w) \in \operatorname{gph} S$ and $DS(v, w)$ is also single-valued such that $\delta w = DS(v, w)[\delta v]$, one infers that S is directionally differentiable at v along δv with the directional derivate $S'(v; \delta v) = \delta w$; see, e.g., [Lev01]. The directional differentiability of the lower-level solution mapping S is asserted in the following theorem.

Theorem 3.4.3 (Directional differentiability)**.** *Let $S : Q_h \to \mathbb{R}^{|\Omega_u|}$ be the solution mapping in Theorem 3.4.1 and (u, h, p) be a feasible solution satisfying the state equation (3.4.2). Then S is directionally differentiable at h along any $\delta h \in T_{Q_h}(h)$. Moreover, the directional derivative $\delta u := S'(h; \delta h)$ is uniquely determined by the following sensitivity equation:*

$$
\left\{\begin{array}{l} D_u F(u, h)\delta u + D_h F(u, h) \delta h + \alpha \nabla^\top \delta p = 0, \\ (\delta u, \alpha \nabla^\top \delta p) \in T_{\operatorname{gph} G}(u, \alpha \nabla^\top p). \end{array}\right. \tag{3.4.5}
$$

Proof. By Theorem 4.1 in [Sha05], the following estimate on the graphical derivative of S holds true:

$$DS(h, u)[\delta h] \subset \left\{ \delta u \in \mathbb{R}^{|\Omega_u|} : 0 \in D_u F(u, h) \delta u + D_h F(u, h) \delta h + DG(u, -F(u, h))[\delta u] \right\}. \tag{3.4.6}$$

With the introduction of the auxiliary variables p and δp such that (u, h, p) satisfies (3.4.2) and $(\delta u, \alpha\nabla^\top \delta p) \in T_{\operatorname{gph} G}(u, \alpha\nabla^\top p)$, the relation (3.4.6) is equivalent to

$$DS(h, u)[\delta h] \subset \Big\{\delta u \in \mathbb{R}^{|\Omega_u|} : (\delta u, \delta h, \delta p) \text{ satisfies the sensitivity equation (3.4.5)}\Big\}. \qquad (3.4.7)$$

Let $\delta h \in T_{Q_h}(h)$ be arbitrarily fixed in the following.

We first show that the set $DS(h, u)[\delta h]$ is nonempty. Following the definition of a tangent cone in (3.2.4), there exists $t^i \to 0^+$, $\delta h^i \to \delta h$ such that $h + t^i\delta h^i \in Q_h$ for all i. Then we have

$$\limsup_{i\to\infty} \frac{\|S(h + t^i\delta h^i) - S(h)\|}{t^i} \leq \kappa\|\delta h\|,$$

where κ is the Lipschitz constant for S near h. As a result, possibly along a subsequence, we have

$$\lim_{i\to\infty} \frac{S(h + t^i\delta h^i) - S(h)}{t^i} = \delta u$$

for some $\delta u \in \mathbb{R}^{|\Omega_u|}$. Thus, we assert that $(\delta h, \delta u) \in T_{\operatorname{gph} S}(h, u)$, or equivalently $\delta u \in DS(h, u)[\delta h]$.

Next we show that δu must be unique among all solutions $(\delta u, \delta p)$ for (3.4.5). Fixing $h \in Q_h$, let $(\delta u^1, \delta p^1)$ and $(\delta u^2, \delta p^2)$ be two solutions for (3.4.5). Then we have

$$D_uF(u, h)(\delta u^1 - \delta u^2) + \alpha\nabla^\top(\delta p^1 - \delta p^2) = 0,$$

which further implies

$$\langle \delta u^1 - \delta u^2, D_uF(u, h)(\delta u^1 - \delta u^2)\rangle + \alpha\langle \nabla\delta u^1 - \nabla\delta u^2, \delta p^1 - \delta p^2\rangle = 0.$$

We claim that $\langle \nabla\delta u^1 - \nabla\delta u^2, \delta p^1 - \delta p^2\rangle \geq 0$. Indeed, we component-wisely distinguish the following three cases.

(1) Consider $j \in \Omega_u$ where $|p_j| < 1$. Then it follows immediately from (3.4.3) that $(\nabla\delta u^1)_j - (\nabla\delta u^2)_j = 0$.

(2) Consider $j \in \Omega_u$ where $(\nabla u)_j \neq 0$. Then from (3.4.3) we have

$$\begin{aligned}
&\langle (\nabla\delta u^1)_j - (\nabla\delta u^2)_j, \delta p_j^1 - \delta p_j^2\rangle \\
&= \langle (\nabla\delta u^1)_j - (\nabla\delta u^2)_j, \frac{1}{|(\nabla u)_j|}(I - p_jp_j^\top)((\nabla\delta u^1)_j - (\nabla\delta u^2)_j)\rangle \\
&\geq \frac{1}{|(\nabla u)_j|}(1 - |p_j|^2)|(\nabla\delta u^1)_j - (\nabla\delta u^2)_j|^2 \geq 0.
\end{aligned}$$

(3) The last case where $j \in \Omega_u$ with $(\nabla u)_j = 0$ and $|p_j| = 1$ further splits into three subcases.

(3a) Consider $(\nabla\delta u^1)_j = 0$, $\langle \delta p_j^1, p_j\rangle \leq 0$ and $(\nabla\delta u^2)_j = 0$, $\langle \delta p_j^2, p_j\rangle \leq 0$. Then as in case (1) we have $(\nabla\delta u^1)_j - (\nabla\delta u^2)_j = 0$.

(3b) Consider $(\nabla\delta u^1)_j = c_1 p_j$ $(c_1 \geq 0)$, $\langle \delta p_j^1, p_j\rangle = 0$ as well as $(\nabla\delta u^2)_j = c_2 p_j$ $(c_2 \geq 0)$, $\langle \delta p_j^2, p_j\rangle = 0$. Then $\langle(\nabla\delta u^1)_j - (\nabla\delta u^2)_j, \delta p_j^1 - \delta p_j^2\rangle = (c_1 - c_2)\langle p_j, \delta p_j^1 - \delta p_j^2\rangle = 0$.

(3c) Consider $(\nabla\delta u^1)_j = 0$, $\langle \delta p_j^1, p_j\rangle \leq 0$ and $(\nabla\delta u^2)_j = cp_j$ $(c \geq 0)$, $\langle \delta p_j^2, p_j\rangle = 0$. Then we have $\langle(\nabla\delta u^1)_j - (\nabla\delta u^2)_j, \delta p_j^1 - \delta p_j^2\rangle = \langle -cp_j, \delta p_j^1 - \delta p_j^2\rangle \geq 0$. The analogous conclusion holds true if we interchange the upper indices ^{1} and ^{2}.

Altogether, our claim is proven. Moreover, since $D_u F(u,h)$ is strictly positive definite, we arrive at $\delta u^1 = \delta u^2$.

Thus, the equality holds in (3.4.7) with both sides being singletons, which concludes the proof. □

Thus, it has been asserted that the solution mapping $S : h \mapsto u(h)$ for the lower-level problem is *B(ouligand)-differentiable* [Rob87], i.e. locally Lipschitz continuous and directionally differentiable, everywhere on Q_h such that, with $\delta u(h;\delta h) = S'(h;\delta h)$, we have

$$u(h+\delta h) = u(h) + \delta u(h;\delta h) + o(\|\delta h\|) \quad \text{as } \delta h \to 0.$$

Furthermore, according to the chain rule, the reduced objective $\widehat{J} : h \to \mathbb{R}$ is also B-differentiable such that

$$\widehat{J}(h+\delta h) = J(u(h),h) + D_h J(u(h),h)\delta h + D_u J(u(h),h)\delta u(h;\delta h) + o(\|\delta h\|) \quad \text{as } \delta h \to 0. \tag{3.4.8}$$

3.5 Stationarity conditions for bilevel optimization

Our bilevel optimization problem (3.3.7) is a special instance of a mathematical program with equilibrium constraints (MPEC). The derivation of appropriate stationarity conditions is a persistent challenge for MPECs; see [LPR96, OKZ98] for more backgrounds on MPECs. Very often, the commonly used constraint qualifications like linear independence constraint qualification (LICQ) or Mangasarian-Fromovitz constraint qualification (MFCQ) are violated for MPECs [YZZ97], and therefore a theoretically sharp and computationally amenable characterization of the variational geometry (such as tangent and normal cones) of the solution set induced by the lower-level problem becomes a major challenge. In this vein, various stationarity concepts are introduced in [SS00] when the lower-level problems are so-called complementarity problems. These stationarity concepts have been further developed and extended during the past decade; see, e.g., [LPR96, OKZ98, Mor06, SS00, Ye05, HK09, HS14]. This research field still remains active in its own right.

In our context of the bilevel optimization problem (3.3.7), it is straightforward to deduce from the expansion formula (3.4.8) that

$$D_h J(u(h),h)\delta h + D_u J(u(h),h)\delta u(h;\delta h) \geq 0 \quad \forall \delta h \in T_{Q_h}(h) \tag{3.5.1}$$

must hold at any local minimizer $(h, u(h))$ for (3.3.7). In fact, condition (3.5.1) is referred to as B(ouligand)-stationarity; see [LPR96]. However, such "primal" stationarity is difficult to realize numerically, since the mapping $\delta h \mapsto \delta u(h; \delta h)$ need not be linear. For this reason, we are motivated to search for stationarity conditions in "primal-dual" form, as they typically appear in the classical KKT conditions for constrained optimization.

Based on the strong regularity condition proven in Theorem 3.4.1 above and the Mordukhovich calculus (see the two-volume monograph [Mor06] for reference), we shall derive the M(ordukhovich)-stationarity for (3.3.7) in Theorem 3.5.1. There the Mordukhovich (or limiting) normal cone of $\operatorname{gph} G$ will appear in the stationarity condition; recall Definition 3.2.4. In particular, one has $N_Q^{(\mathrm{M})}(\cdot) = N_Q(\cdot)$ whenever Q is convex. Following (3.4.3) and (3.4.4), the Mordukhovich normal cone of $\operatorname{gph} G$ can be calculated as:

$$N^{(\mathrm{M})}_{\operatorname{gph} G}(u, \alpha\nabla^\top p) = \Bigg\{ (\alpha\nabla^\top w, -v) : \\ \left\{\begin{array}{ll} w_j = \xi_j - \langle \xi_j, p_j\rangle p_j,\ (\nabla v)_j = |(\nabla u)_j|\xi_j\ (\xi_j \in \mathbb{R}^2) & \text{if } (\nabla u)_j \neq 0, \\ w_j \in \mathbb{R}^2,\ (\nabla v)_j = 0 & \text{if } |p_j| < 1, \\ w_j \in \mathbb{R}^2,\ (\nabla v)_j = 0, \text{ or} & \\ \langle w_j, p_j\rangle = 0,\ (\nabla v)_j = c p_j\ (c \in \mathbb{R}), \text{ or} & \\ \langle w_j, p_j\rangle \leq 0,\ (\nabla v)_j = c p_j\ (c \leq 0) & \text{if } (\nabla u)_j = 0,\ |p_j| = 1. \end{array}\right. \Bigg\}. \tag{3.5.2}$$

We are now ready to present the M-stationarity condition for (3.3.7). Given that the strong regularity condition is satisfied at any feasible solution (u, h, p) as justified in Theorem 3.4.1, M-stationarity of a local minimizer for (3.3.7) follows as a direct consequence of Theorem 3.2.16. Notably, the strong regularity condition serves as a proper constraint qualification in deriving the M-stationarity.

Theorem 3.5.1 (M-stationarity). *Let $(u, h, p) \in \mathbb{R}^{|\Omega_u|} \times Q_h \times Q_p$ be any feasible point satisfying (3.4.2). If (u, h) is a local minimizer for the bilevel optimization problem (3.3.7), then the following M-stationarity condition must hold true for some $(w, v) \in \left(\mathbb{R}^{|\Omega_u|}\right)^2 \times \mathbb{R}^{|\Omega_u|}$:*

$$\left\{\begin{array}{l} D_u J(u,h)^\top + \alpha\nabla^\top w + D_u F(u,h)^\top v = 0, \\ 0 \in D_h J(u,h)^\top + D_h F(u,h)^\top v + N_{Q_h}(h), \\ (\alpha\nabla^\top w, -v) \in N^{(\mathrm{M})}_{\operatorname{gph} G}(u, \alpha\nabla^\top p), \end{array}\right. \tag{3.5.3}$$

where $N^{(\mathrm{M})}_{\operatorname{gph} G}$ is the Mordukhovich normal cone of $\operatorname{gph} G$ *given in (3.5.2).*

Though theoretically sharp, the M-stationarity condition in the above theorem is in general not guaranteed by numerical algorithms. Instead, we resort to a Clarke-type stationarity, termed C-stationarity in the following corollary. The C-stationarity is slightly weaker than the M-stationarity due to the relation $N^{(\mathrm{M})}_{\operatorname{gph} G}(u, \alpha\nabla^\top p) \subset N^{(\mathrm{C})}_{\operatorname{gph} G}(u, \alpha\nabla^\top p)$, but can be guaranteed by a projected-gradient-type algorithm proposed in section 3.6 below.

Corollary 3.5.2 (C-stationarity)**.** *Let $(u,h,p) \in \mathbb{R}^{|\Omega_u|} \times Q_h \times Q_p$ be any feasible point satisfying (3.4.2). If (u,h) is a local minimizer for the bilevel optimization problem (3.3.7), the following C-stationarity condition must hold true for some $(w,v) \in \left(\mathbb{R}^{|\Omega_u|}\right)^2 \times \mathbb{R}^{|\Omega_u|}$:*

$$\begin{cases} D_u J(u,h)^\top + \alpha\nabla^\top w + D_u F(u,h)^\top v = 0, \\ 0 \in D_h J(u,h)^\top + D_h F(u,h)^\top v + N_{Q_h}(h), \\ (\alpha\nabla^\top w, -v) \in N^{(\mathrm{C})}_{\operatorname{gph} G}(u, \alpha\nabla^\top p), \end{cases} \tag{3.5.4}$$

where

$$N^{(\mathrm{C})}_{\operatorname{gph} G}(u, \alpha\nabla^\top p) = \Bigg\{ (\alpha\nabla^\top w, -v) : \\ \begin{cases} w_j = \xi_j - \langle \xi_j, p_j \rangle p_j, \ (\nabla v)_j = |(\nabla u)_j| \xi_j \ (\xi_j \in \mathbb{R}^2) & \text{if } (\nabla u)_j \neq 0, \\ w_j \in \mathbb{R}^2, \ (\nabla v)_j = 0 & \text{if } |p_j| < 1, \\ (\nabla v)_j = c p_j \ (c \in \mathbb{R}), \ \langle w_j, (\nabla v)_j \rangle \geq 0 & \text{if } (\nabla u)_j = 0, \ |p_j| = 1. \end{cases} \Bigg\}. \tag{3.5.5}$$

We say that *strict complementarity* holds at a feasible point (u,h,p) whenever the biactive set is empty, i.e.

$$\{ j \in \Omega_u : (\nabla u)_j = 0, \ |p_j| = 1 \} = \emptyset. \tag{3.5.6}$$

Under strict complementarity, one immediately observes the equivalence of M- and C-stationarity as $N^{(\mathrm{M})}_{\operatorname{gph} G}(u, \alpha\nabla^\top p) = N^{(\mathrm{C})}_{\operatorname{gph} G}(u, \alpha\nabla^\top p)$. The scenarios of strict complementarity are studied in detail in section 3.6.1, where it will become evident to the reader that all B-, M-, and C-stationarity concepts are equivalent under the strict complementarity; see Corollary 3.6.3.

3.6 Hybrid projected gradient method

This section is devoted to the development and the convergence analysis of a hybrid projected gradient algorithm to compute a C-stationary point for the bilevel optimization problem (3.3.7). Most existing numerical solvers for MPECs adopt regularization/smoothing/relaxation on the complementary structure in the lower-level problem, see e.g. [FLP98, Sch01, FLRS06], even though the complementary structure induced by (3.4.2) is more involved than those in the previous works due to the presence of nonlinearity. Motivated by the recent work in [HS14], here we devise an algorithm which avoids redundant regularization, e.g., when the current iterate is a continuously differentiable point for the reduced objective $\widehat{J}$.

3.6.1 Differentiability given strict complementarity

In this subsection, we assume that strict complementarity, i.e. condition (3.5.6), holds at a feasible point (u,h,p). In this scenario, the sensitivity equation (3.4.5) is fully characterized by the following linear system:

$$\begin{bmatrix} D_u F(u,h) & \alpha\nabla^\top \\ (-I + pp^\top)\nabla & \operatorname{diag}(|\nabla u| e) \end{bmatrix} \begin{bmatrix} \delta u \\ \delta p \end{bmatrix} = \begin{bmatrix} -D_h F(u,h)\delta h \\ 0 \end{bmatrix}. \tag{3.6.1}$$

Here e is the identity vector in $\left(\mathbb{R}^{|\Omega|}\right)^2$, i.e. $e_j = (1,1)$ for all $j \in \Omega_u$, and $\operatorname{diag}(|\nabla u|e)$ denotes a diagonal matrix with its diagonal elements given by the vector $|\nabla u|e$. As a special case in Theorem 3.4.3, for any given $\delta h \in T_{Q_h}(h)$, the linear system (3.6.1) always admits a solution $(\delta u, \delta p)$ which is unique in δu. Thus, the differential mapping $\frac{\delta u}{\delta h}(h) : \delta h \mapsto \delta u$ defined by equation (3.6.1) is a continuous linear mapping, and therefore the reduced objective $\widehat{J}$ in (3.4.1) is continuously differentiable at h. On the other hand, the adjoint of the differential $\frac{\delta u}{\delta h}(h)$, denoted by $\frac{\delta u}{\delta h}(h)^\top$, is required when computing $D_h\widehat{J}(h)$. This will be addressed through the adjoint equation in Theorem 3.6.2 below.

Lemma 3.6.1. *Assume that (u,h,p) is a feasible point satisfying (3.4.2) and strict complementarity holds at (u,h,p). Let $\Pi_{\delta u}$ be a canonical projection such that $\Pi_{\delta u}(\delta u, \delta p) = (\delta u, 0)$ for all $(\delta u, \delta p) \in \mathbb{R}^{|\Omega_u|} \times \left(\mathbb{R}^{|\Omega_u|}\right)^2$. Then the following relations hold true:*

(i) $\operatorname{Ker}\begin{bmatrix} D_uF(u,h)^\top & \nabla^\top(-I+pp^\top) \\ \alpha\nabla & \operatorname{diag}(|\nabla u|e) \end{bmatrix} \subset \operatorname{Ker}\Pi_{\delta u}$.

(ii) $\operatorname{Ran}\Pi_{\delta u} \subset \operatorname{Ran}\begin{bmatrix} D_uF(u,h)^\top & \nabla^\top(-I+pp^\top) \\ \alpha\nabla & \operatorname{diag}(|\nabla u|e) \end{bmatrix}$.

Proof. We first prove (i). For this purpose, let

$$\begin{bmatrix} D_uF(u,h)^\top & \nabla^\top(-I+pp^\top) \\ \alpha\nabla & \operatorname{diag}(|\nabla u|e) \end{bmatrix}\begin{bmatrix} v \\ \eta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which implies

$$\begin{aligned} 0 &= \langle v, D_uF(u,h)^\top v\rangle + \langle \nabla v, (-I+pp^\top)\eta\rangle \\ &= \langle v, D_uF(u,h)^\top v\rangle + \frac{1}{\alpha}\langle |\nabla u|\eta, (I-pp^\top)\eta\rangle \\ &= \langle v, D_uF(u,h)^\top v\rangle + \frac{1}{\alpha}\sum_{j\in\Omega_u} |(\nabla u)_j|(|\eta_j|^2 - |\langle p_j, \eta_j\rangle|^2). \end{aligned}$$

Owing to the strict positive definiteness of $D_uF(u,h)$ as well as the non-negativity of the second term in the above equation, we verify that $v = 0$.

To justify (ii), in view of the fundamental theorem of linear algebra, it suffices to prove

$$\operatorname{Ker}\begin{bmatrix} D_uF(u,h) & \alpha\nabla^\top \\ (-I+pp^\top)\nabla & \operatorname{diag}(|\nabla u|e) \end{bmatrix} \subset \operatorname{Ker}\Pi_{\delta u}.$$

For this purpose, consider

$$\begin{bmatrix} D_uF(u,h) & \alpha\nabla^\top \\ (-I+pp^\top)\nabla & \operatorname{diag}(|\nabla u|e) \end{bmatrix}\begin{bmatrix} \delta u \\ \delta p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{3.6.2}$$

Then we have

$$\langle \delta u, D_uF(u,h)\delta u\rangle + \alpha\langle \delta p, pp^\top\nabla\delta u\rangle + \alpha\langle \delta p, |\nabla u|\delta p\rangle = 0. \tag{3.6.3}$$

Due to the strict complementarity, only two possible scenarios may occur. If $(\nabla u)_j \neq 0$, then the second row of equation (3.6.2) yields $\delta p_j = \frac{1}{|(\nabla u)_j|}(I - p_jp_j^\top)(\nabla\delta u)_j$, and thus $\langle \delta p_j, p_jp_j^\top(\nabla\delta u)_j\rangle \geq$

0. If $|p_j| < 1$, then $(\nabla u)_j = 0$ and $0 = |(I - p_j p_j^\top)(\nabla \delta u)_j| \geq (1 - |p_j|^2)|(\nabla \delta u)_j|$, which implies $\langle \delta p_j, p_j p_j^\top (\nabla \delta u)_j \rangle = 0$. Altogether, we have shown $\langle \delta p, pp^\top \nabla \delta u \rangle \geq 0$. Moreover, since the third term in (3.6.3) is also non-negative and $D_u F(u,h) = -\mu\Delta + K(h)^\top K(h)$ is strictly positive definite, we must have $\delta u = 0$. Thus, (ii) is proven. □

Theorem 3.6.2. *As in Lemma 3.6.1, assume that (u,h,p) is a feasible point satisfying (3.4.2) and strict complementarity holds at (u,h,p). Then $\frac{\delta u}{\delta h}(h)^\top$ is a linear mapping such that $\frac{\delta u}{\delta h}(h)^\top : \zeta \mapsto D_h F(u,h)^\top v$ with $(\zeta, v, \eta) \in \mathbb{R}^{|\Omega_u|} \times \mathbb{R}^{|\Omega_u|} \times \left(\mathbb{R}^{|\Omega_u|}\right)^2$ satisfying the following adjoint equation:*

$$\begin{bmatrix} D_u F(u,h)^\top & \nabla^\top(-I + pp^\top) \\ \alpha\nabla & \operatorname{diag}(|\nabla u| e) \end{bmatrix} \begin{bmatrix} v \\ \eta \end{bmatrix} = \begin{bmatrix} -\zeta \\ 0 \end{bmatrix}. \tag{3.6.4}$$

Proof. It follows from Lemma 3.6.1 that $\zeta \mapsto v$ is a continuous linear mapping and, therefore, so is $\frac{\delta u}{\delta h}(h)^\top$. To show the adjoint relation between $\frac{\delta u}{\delta h}(h)$ and $\frac{\delta u}{\delta h}(h)^\top$, consider an arbitrary pair $(\delta u, \delta h, \delta p)$ which satisfies (3.6.1), i.e. $\delta u = \frac{\delta u}{\delta h}(h)\delta h$, and (ζ, v, η) which satisfies (3.6.4). Then we derive that

$$\begin{aligned} \left\langle \zeta, \frac{\delta u}{\delta h}(h)\delta h \right\rangle &= -\left\langle \begin{bmatrix} \delta u \\ \delta p \end{bmatrix}, \begin{bmatrix} D_u F(u,h)^\top & \nabla^\top(-I + pp^\top) \\ \alpha\nabla & \operatorname{diag}(|\nabla u| e) \end{bmatrix} \begin{bmatrix} v \\ \eta \end{bmatrix} \right\rangle \\ &= -\left\langle \begin{bmatrix} D_u F(u,h) & \alpha\nabla^\top \\ (-I + pp^\top)\nabla & \operatorname{diag}(|\nabla u| e) \end{bmatrix} \begin{bmatrix} \delta u \\ \delta p \end{bmatrix}, \begin{bmatrix} v \\ \eta \end{bmatrix} \right\rangle \\ &= \langle v, D_h F(u,h)\delta h \rangle = \langle D_h F(u,h)^\top v, \delta h \rangle = \left\langle \frac{\delta u}{\delta h}(h)^\top \zeta, \delta h \right\rangle, \end{aligned}$$

which concludes the proof. □

As a consequence of Theorem 3.6.2, at a feasible point (u,h,p) where the strict complementarity holds, the gradient of the reduced objective can be calculated as

$$D_h \widehat{J}(h)^\top = D_h J(u,h)^\top + \frac{\delta u}{\delta h}(h)^\top D_u J(u,h)^\top = D_h J(u,h)^\top + D_h F(u,h)^\top v, \tag{3.6.5}$$

where (v, η) satisfies the adjoint equation (3.6.4) with $\zeta = D_u J(u,h)^\top$. For the sake of our convergence analysis in section 3.6.3, we also introduce an auxiliary variable w defined by

$$w := \frac{1}{\alpha}(-I + p(p)^\top)\eta, \tag{3.6.6}$$

which parallels the auxiliary variable w^γ later in (3.6.16) for the smoothing case. To conclude section 3.6.1, we point out that one can readily deduce from (3.6.5) the equivalence among the B-, M-, and C-stationarity under strict complementarity.

Corollary 3.6.3 (Stationarity under strict complementarity)**.** *If strict complementarity holds at a feasible point (u,h,p), then B-stationarity (3.5.1), M-stationarity (3.5.3), and C-stationarity (3.5.4) are all equivalent.*

3.6.2 Local smoothing at a non-differentiable point

The solution mapping $h \mapsto u$ for the lower-problem in (3.3.7) is only B-differentiable (rather than continuously differentiable) at a feasible point (u, h, p) where the biactive set $\{j \in \Omega_u : (\nabla u)_j = 0,\ |p_j| = 1\}$ is nonempty. In this scenario, continuous optimization techniques are not directly applicable. Instead, we utilize a local smoothing approach by replacing the Lipschitz continuous function $\|\cdot\|_1$ in (3.3.2) by a C^2-approximation $\|\cdot\|_{1,\gamma} : \left(\mathbb{R}^{|\Omega_u|}\right)^2 \to \mathbb{R}$, which is defined for each $\gamma > 0$ by $\|p\|_{1,\gamma} := \sum_{j\in\Omega_u} \varphi_\gamma(p_j)$ with

$$\varphi_\gamma(s) = \begin{cases} -\frac{1}{8\gamma^3}|s|^4 + \frac{3}{4\gamma}|s|^2 & \text{if } |s| < \gamma, \\ |s| - \frac{3\gamma}{8} & \text{if } |s| \geq \gamma. \end{cases} \tag{3.6.7}$$

The first-order and second-order derivatives of φ_γ can be calculated as

$$\varphi'_\gamma(s) = \begin{cases} (\frac{3}{2\gamma} - \frac{1}{2\gamma^3}|s|^2)s & \text{if } |s| < \gamma, \\ \frac{1}{|s|}s & \text{if } |s| \geq \gamma. \end{cases} \tag{3.6.8}$$

and

$$\varphi''(s) = \begin{cases} (\frac{3}{2\gamma} - \frac{1}{2\gamma^3}|s|^2)I_{\mathbb{R}^2} - \frac{1}{\gamma^3}ss^\top & \text{if } |s| < \gamma, \\ \frac{1}{|s|}I_{\mathbb{R}^2} - \frac{1}{|s|^3}ss^\top & \text{if } |s| \geq \gamma. \end{cases} \tag{3.6.9}$$

We remark that the same smoothing function was used in [KP13] for parameter learning, but other choices are possible as well.

The resulting smoothed bilevel optimization problem appears as

$$\begin{array}{ll} \min & J(u^\gamma, h) \\ \text{s.t.} & u^\gamma = \arg\min_u \dfrac{\mu}{2}\|\nabla u\|^2 + \dfrac{1}{2}\|K(h)u - z\|^2 + \alpha\|\nabla u\|_{1,\gamma}, \\ & u^\gamma \in \mathbb{R}^{|\Omega_u|},\ h \in Q_h. \end{array} \tag{3.6.10}$$

The corresponding Euler-Lagrange equation for the lower-level problem in (3.6.10) is given by

$$r(u^\gamma; h, \gamma) := (-\mu\Delta + K(h)^\top K(h))u^\gamma - K(h)^\top z + \alpha\nabla^\top(\varphi'_\gamma(\nabla u^\gamma)) = 0, \tag{3.6.11}$$

which induces a continuously differentiable mapping $h \mapsto u^\gamma(h)$ according to the (classical) implicit function theorem. Moreover, the sensitivity equation for (3.6.11) is given by

$$\left(D_u F(u^\gamma, h) + \alpha\nabla^\top\varphi''_\gamma(\nabla u^\gamma)\nabla\right) D_h u^\gamma(h) = -D_h F(u^\gamma, h). \tag{3.6.12}$$

Analogous to (3.4.1), we may also reformulate the smoothed bilevel problem (3.6.10) in the reduced form as

$$\begin{array}{ll} \min & \breve{J}_\gamma(h) := J(u^\gamma(h), h) \\ \text{s.t.} & h \in Q_h. \end{array} \tag{3.6.13}$$

The gradient of $\breve{J}_\gamma$ can be calculated as

$$D_h\breve{J}_\gamma(h)^\top = D_h J(u^\gamma, h)^\top + D_h F(u^\gamma, h)^\top v^\gamma, \tag{3.6.14}$$

where v^γ satisfies the adjoint equation

$$\Big(D_u F(u^\gamma,h)^\top + \alpha\nabla^\top \varphi''_\gamma(\nabla u^\gamma)\nabla\Big)\, v^\gamma = -D_u J(u^\gamma,h)^\top. \tag{3.6.15}$$

Thus, any stationary point (u^γ,h) of the smoothed bilevel optimization problem (3.6.10) must satisfy the following stationarity condition

$$\begin{cases} F(u^\gamma,h)+\alpha\nabla^\top p^\gamma = 0,\\ p^\gamma = \varphi'_\gamma(\nabla u^\gamma),\\ D_u F(u^\gamma,h)^\top v^\gamma + \alpha\nabla^\top w^\gamma = -D_u J(u^\gamma,h)^\top,\\ w^\gamma = \varphi''_\gamma(\nabla u^\gamma)\nabla v^\gamma,\\ 0\in D_h J(u^\gamma,h)^\top + D_h F(u^\gamma,h)^\top v^\gamma + N_{Q_h}(h), \end{cases} \tag{3.6.16}$$

for some $p^\gamma \in \left(\mathbb{R}^{|\Omega_u|}\right)^2$, $w^\gamma \in \left(\mathbb{R}^{|\Omega_u|}\right)^2$, and $v^\gamma \in \mathbb{R}^{|\Omega_u|}$.

We remark that finding a stationary point of the (smooth) constrained minimization problem (3.6.13) can be accomplished by standard optimization algorithms; see [NW06]. As a subroutine in Algorithm 3.6.5 below, we adopt a simple projected gradient method whose convergence analysis can be found in [GB82]. The following theorem establishes the consistency on how a stationary point of the smoothed bilevel problem (3.6.10) approaches a C-stationary point of the original bilevel problem (3.3.7) as γ vanishes.

Theorem 3.6.4 (Consistency of smoothing)**.** *Let $\{\gamma^k\}$ be any sequence of positive scalars such that $\gamma^k \to 0^+$. For each γ^k, let $(u^k,h^k)\in\mathbb{R}^{|\Omega_u|}\times Q_h$ be a stationary point of (3.6.10) such that condition (3.6.16) holds for some $(p^k,w^k,v^k)\in\left(\mathbb{R}^{|\Omega_u|}\right)^2\times\left(\mathbb{R}^{|\Omega_u|}\right)^2\times\mathbb{R}^{|\Omega_u|}$. Then any accumulation point of $\{(u^k,h^k,p^k,w^k,v^k)\}$ is a feasible C-stationary point for (3.3.7) satisfying (3.4.2) and (3.5.4).*

Proof. Let (u^*,h^*,p^*,w^*,v^*) be an arbitrary accumulation point of $\{(u^k,h^k,p^k,w^k,v^k)\}$. Then the first condition in (3.4.2) and the first condition in (3.5.4) immediately follow from continuity. The second condition in (3.5.4) also follows due to the closedness of the normal cone mapping $N_{Q_h}(\cdot)$; see, e.g., Proposition 6.6 in [RW98].

For those $j\in\Omega_u$ where $(\nabla u^*)_j\neq 0$, we have for all sufficiently large k that $p_j^k = \frac{(\nabla u^k)_j}{|(\nabla u^k)_j|}$, and therefore $p_j^* = \frac{(\nabla u^*)_j}{|(\nabla u^*)_j|}$. On the other hand, $p_j^*\in Q_p$ clearly holds if $(\nabla u^*)_j = 0$. Altogether, the feasibility of (u^*,h^*,p^*) is verified.

It remains to show $(\alpha\nabla^\top w^*, -v^*)\in N^{(\mathrm{C})}_{\operatorname{gph} G}(u^*,\alpha\nabla^\top p^*)$, for which the proof is divided into three cases as follows.

(1) If $(\nabla u^*)_j\neq 0$, then we have for all sufficiently large k that $|(\nabla u^k)_j|\geq\gamma^k$ and therefore

$$w_j^k = \frac{1}{|(\nabla u^k)_j|}(\nabla v^k)_j - \frac{1}{|(\nabla u^k)_j|}\langle(\nabla v^k)_j, p_j^k\rangle p_j^k.$$

Passing $k\to\infty$, the first condition in (3.5.5) is fulfilled with $\xi_j = \frac{1}{|(\nabla u^*)_j|}(\nabla v^*)_j$.

(2) If $|p_j^*| < 1$, then we have for all sufficiently large k that $|p_j^k| < 1$ and, therefore, $|(\nabla u^k)_j| < \gamma^k$. This implies $(\nabla u^*)_j = 0$. Let $q_j \in \mathbb{R}^2$ be an arbitrary accumulation point of the uniformly bounded sequence $\{(\nabla u^k)_j/\gamma^k\}$. We obviously have $|q_j| \le 1$. Then it follows from $p^k = \varphi'_{\gamma^k}(\nabla u^k)$ that $p_j^* = (3/2 - |q_j|^2/2)q_j$. Since $|p_j^*| < 1$, we must have $|q_j| < 1$. Since $w^k = \varphi''_{\gamma^k}(\nabla u^k)\nabla v^k$, we have

$$\gamma^k w_j^k = \left(\frac{3}{2} - \frac{|(\nabla u^k)_j|^2}{2(\gamma^k)^2}\right)(\nabla v^k)_j - \left\langle (\nabla v^k)_j, \frac{(\nabla u^k)_j}{\gamma^k}\right\rangle \frac{(\nabla u^k)_j}{\gamma^k}.$$

Passing $k \to \infty$, we obtain

$$\frac{3 - |q_j|^2}{2}(\nabla v^*)_j - \langle q_j, (\nabla v^*)_j\rangle q_j = 0,$$

which indicates that $(\nabla v^*)_j = c q_j$ for some $c \in \mathbb{R}$. Thus it follows that $\frac{3}{2}(1 - |q_j|^2)(\nabla v^*)_j = 0$, and thus $(\nabla v^*)_j = 0$ as requested by the second condition in (3.5.5).

(3) Now we investigate the third condition in (3.5.5) where $(\nabla u^*)_j = 0$ and $|p_j^*| = 1$ under the following two circumstances.

(3a) There exists an infinite index subset $\{k'\} \subset \{k\}$ such that $(\nabla u^{k'})_j \ge \gamma^{k'}$ for all k'. Then it holds for all k' that

$$\begin{cases} |p_j^{k'}| = 1, \\ w_j^{k'} = \dfrac{1}{|(\nabla u^{k'})_j|}(\nabla v^{k'})_j - \dfrac{1}{|(\nabla u^{k'})_j|}\langle (\nabla v^{k'})_j, p_j^{k'}\rangle p_j^{k'}, \end{cases} \tag{3.6.17}$$

and therefore

$$\begin{cases} \langle w_j^{k'}, p_j^{k'}\rangle = 0, \\ |(\nabla u^{k'})_j| w_j^{k'} = (\nabla v^{k'})_j - \langle (\nabla v^{k'})_j, p_j^{k'}\rangle p_j^{k'}. \end{cases}$$

Passing $k' \to \infty$, we have $\langle w_j^*, p_j^*\rangle = 0$ and $(\nabla v^*)_j - \langle (\nabla v^*)_j, p_j^*\rangle p_j^* = 0$. Thus the third condition in (3.5.5) is fulfilled.

(3b) There exists an infinite index subset $\{k'\} \subset \{k\}$ such that $(\nabla u^{k'})_j < \gamma^{k'}$ for all k'. Then analogous to case (2), we have for all k' that

$$\begin{cases} p_j^{k'} = \left(\dfrac{3}{2\gamma^{k'}} - \dfrac{|\nabla u_j^{k'}|^2}{2(\gamma^{k'})^3}\right)\nabla u_j^{k'}, \\ \gamma^{k'} w_j^{k'} = \left(\dfrac{3}{2} - \dfrac{|(\nabla u^{k'})_j|^2}{2(\gamma^{k'})^2}\right)(\nabla v^{k'})_j - \left\langle (\nabla v^{k'})_j, \dfrac{(\nabla u^{k'})_j}{\gamma^{k'}}\right\rangle \dfrac{(\nabla u^{k'})_j}{\gamma^{k'}}. \end{cases} \tag{3.6.18}$$

Let $q_j \in \mathbb{R}^2$ be an arbitrary accumulation point of the uniformly bounded sequence $\{(\nabla u^{k'})_j/\gamma^{k'}\}$. Then we have $p_j^* = (\frac{3}{2} - \frac{1}{2}|q_j|^2)q_j$. It follows from $|p_j^*| = 1$ that $|q_j| = 1$ must hold. Since this holds true for an arbitrary accumulation point q_j, we infer that $\lim_{k'\to\infty}(\nabla u^{k'})_j/\gamma^k = p_j^*$ and further from the second equation in (3.6.18) that $(\nabla v^*)_j - \langle (\nabla v^*)_j, p_j^* \rangle p_j^* = 0$, i.e. $(\nabla v^*)_j = cp_j^*$ for some $c \in \mathbb{R}$. On the other hand, equation (3.6.18) also yields that

$$\begin{aligned}\langle w_j^{k'}, (\nabla v^{k'})_j \rangle &= \left(\frac{3}{2\gamma^{k'}} - \frac{|(\nabla u^{k'})_j|^2}{2(\gamma^{k'})^3} \right) |(\nabla v^{k'})_j|^2 - \frac{1}{\gamma^{k'}} \left| \left\langle (\nabla v^{k'})_j, \frac{(\nabla u^{k'})_j}{\gamma^{k'}} \right\rangle \right|^2 \\ &\geq \frac{3}{2\gamma^{k'}} \left(1 - \left| \frac{(\nabla u^{k'})_j}{\gamma^{k'}} \right|^2 \right) |(\nabla v^{k'})_j|^2 \geq 0.\end{aligned}$$

Passing $k' \to \infty$, the third condition in (3.5.5) is again fulfilled.

□

3.6.3 Hybrid projected gradient method

Now we present a hybrid projected gradient method for finding a C-stationary point of the bilevel optimization problem (3.3.7). In a nutshell, at a feasible point (u^k, h^k, p^k) where the strict complementarity holds, we calculate $D_h \widehat{J}(h^k)^\top$ according to formula (3.6.5) and perform a projected gradient step by setting

$$\widehat{h}^k(\tau^k) := \mathrm{Proj}_{Q_h}[h^k - \tau^k D_h \widehat{J}(h^k)^\top] \tag{3.6.19}$$

for some proper step size $\tau^k > 0$. If the strict complementarity is violated at (u^k, h^k, p^k), we rather perform a projected gradient step on the smoothed bilevel problem (3.6.10) with $\gamma = \gamma^k > 0$, i.e.

$$\breve{h}^k(\tau^k) := \mathrm{Proj}_{Q_h}[h^k - \tau^k D_h \breve{J}_{\gamma^k}(h^k)^\top]. \tag{3.6.20}$$

In addition, we are cautioned against a critical case where the step size τ^k in (3.6.19) tends to zero along the iterations. This case may possibly occur, provided that the $\{(u^k, h^k, p^k)\}$ converges to some $\{(u^*, h^*, p^*)\}$ where the strict complementarity breaks, even if the strict complementarity holds for each feasible point (u^k, h^k, p^k). In such a critical case, we also resort to the smoothed bilevel problem as in (3.6.20). The overall hybrid algorithm is detailed below.

Algorithm 3.6.5 (Hybrid projected gradient method)**.**

```
Require: inputs $\alpha > 0$, $0 \leq \mu \ll \alpha$, $0 < \underline{\tau} \ll \bar{\tau}$, $0 < \sigma_J < 1$, $0 < \rho_\tau < 1$, $0 < \rho_\gamma < 1$, $\sigma_h > 0$, $\mathrm{tol}_h > 0$, $\mathrm{tol}_\gamma > 0$.
1: Initialize $\gamma^1 > 0$, a feasible point $(u^1, h^1, p^1) \in \mathbb{R}^{|\Omega_u|} \times Q_h \times (\mathbb{R}^{|\Omega_u|})^2$ satisfying (3.4.2), $\widetilde{u}^1 := u^1$, $\widetilde{p}^1 := p^1$, $\mathcal{I} := \{1\}$, and $k := 1$.
2: loop
3:     if the strict complementarity condition (3.5.6) is violated at $(\widetilde{u}^k, h^k, \widetilde{p}^k)$ (i.e. the biactive set $\{j \in \Omega_u : (\nabla \widetilde{u}^k)_j = 0, |\widetilde{p}_j^k| = 1\}$ is nonempty) or $J(\widetilde{u}^k, h^k) > J(u^{\max(\mathcal{I})}, h^{\max(\mathcal{I})})$ then
4:         Go to step 16.
```

```
5:    end if
6:    Set $u^k := \widetilde{u}^k$, $p^k := \widetilde{p}^k$. Compute $D_h\widehat{J}(h^k)^\top$ using formula (3.6.5). Define $\widehat{h}^k(\cdot)$ as in (3.6.19).
7:    if $\|\widehat{h}^k(\bar{\tau}) - h^k\| \leq \mathrm{tol}_h$ then
8:      Return $(u^k, h^k)$ as a C-stationary point of (3.3.7) and terminate the algorithm.
9:    end if
10:   Perform the backtracking line search on $\widehat{h}^k(\cdot)$, i.e. find $\tau^k$ as the largest element in $\{\bar{\tau}(\rho_\tau)^l : l = 0, 1, 2, ...\}$ such that $\widehat{h}^k(\tau^k)$ fulfills the following Armijo-type condition:
          $$\widehat{J}(\widehat{h}^k(\tau^k)) \leq \widehat{J}(h^k) + \sigma_J D_h\widehat{J}(h^k)(\widehat{h}^k(\tau^k) - h^k). \tag{3.6.21}$$
11:   if $\tau^k < \underline{\tau}$ then
12:     Go to step 16.
13:   end if
14:   Set $h^{k+1} := \widehat{h}^k(\tau^k)$ and $\mathcal{I} := \mathcal{I} \cup \{k\}$. Generate $\widetilde{u}^{k+1} \in \mathbb{R}^{|\Omega_u|}$ and $\widetilde{p}^{k+1} \in \left(\mathbb{R}^{|\Omega_u|}\right)^2$ such that $(\widetilde{u}^{k+1}, h^{k+1}, \widetilde{p}^{k+1})$ satisfies the state equation (3.4.2).
15:   Set $\gamma^{k+1} := \gamma^k$. Go to step 26.
16:   Solve equation (3.6.11) with $(\gamma, h) = (\gamma^k, h^k)$ for $u^\gamma =: u^k$, and equation (3.6.15) with $(\gamma, u^\gamma, h) = (\gamma^k, u^k, h^k)$ for $v^\gamma =: v^k$. Then calculate $D_h\breve{J}_{\gamma^k}(h^k)^\top$ using formula (3.6.14). Define $\breve{h}^k(\cdot)$ as in (3.6.20).
17:   if $\|\breve{h}^k(\bar{\tau}) - h^k\| \leq \sigma_h\gamma^k$ then
18:     if $\gamma^k = \mathrm{tol}_\gamma$ then
19:       Return $(u^k, h^k)$ as a C-stationary point of (3.3.7) and terminate the algorithm.
20:     else
21:       Set $\gamma^{k+1} := \max(\rho_\gamma\gamma^k, \mathrm{tol}_\gamma)$ and $(\widetilde{u}^{k+1}, h^{k+1}, \widetilde{p}^{k+1}) := (\widetilde{u}^k, h^k, \widetilde{p}^k)$. Go to step 26.
22:     end if
23:   end if
24:   Perform the backtracking line search on $\breve{h}^k(\cdot)$, i.e. find $\tau^k$ as the largest element in $\{\bar{\tau}(\rho_\tau)^l : l = 0, 1, 2, ...\}$ such that $\breve{h}^k(\tau^k)$ fulfills the following Armijo-type condition:
          $$\breve{J}_{\gamma^k}(\breve{h}^k(\tau^k)) \leq \breve{J}_{\gamma^k}(h^k) + \sigma_J D_h\breve{J}_{\gamma^k}(h^k)(\breve{h}^k(\tau^k) - h^k). \tag{3.6.22}$$
25:   Set $h^{k+1} := \breve{h}^k(\tau^k)$. Generate $\widetilde{u}^{k+1} \in \mathbb{R}^{|\Omega_u|}$ and $\widetilde{p}^{k+1} \in \left(\mathbb{R}^{|\Omega_u|}\right)^2$ such that $(\widetilde{u}^{k+1}, h^{k+1}, \widetilde{p}^{k+1})$ satisfies the state equation (3.4.2). Set $\gamma^{k+1} := \gamma^k$.
26:   Set $k := k + 1$.
27: end loop
```

In the following, we prove convergence of Algorithm 3.6.5 towards C-stationarity. To begin with, we collect a technical result from Lemma 3 in [GB82], which will be used several times in our convergence analysis.

Lemma 3.6.6. *The mappings $\tau^k \mapsto \|\widehat{h}^k(\tau^k) - h^k\|/\tau^k$ and $\tau^k \mapsto \|\breve{h}^k(\tau^k) - h^k\|/\tau^k$ are both monotonically decreasing on* $[0, \infty)$.

Based on Lemma 3.6.6, it is shown in the following lemma that the backtracking line searches in Algorithm 3.6.5 enjoy good properties.

Lemma 3.6.7. *The backtracking line searches in steps 10 and 24 of Algorithm 3.6.5 always terminate with success after finitely many trails.*

Proof. As the line search in step 24 is performed on the continuously differentiable objective $\breve{J}_{\gamma^k}$, the proof of Proposition 2 in [GB82] can be directly applied.

However, this proof needs to be adapted for step 10 since it is performed on the B-differentiable objective $\widehat{J}$. In this case, we proceed with a proof by contradiction. Assume that (3.6.21) is violated for all $\tau^k = \tau_l^k := \bar{\tau}(\rho_\tau)^l$ with $l = 0, 1, 2, ...$ Then h^k cannot be stationary, since otherwise $\widehat{h}^k(\tau^k) = h^k$ and (3.6.21) holds true for any $\tau^k > 0$.

Since $\widehat{J}$ is B-differentiable, we have

$$\widehat{J}(\widehat{h}^k(\tau_l^k)) - \widehat{J}(h^k) = D_h\widehat{J}(h^k)(\widehat{h}^k(\tau_l^k) - h^k) + o(\|\widehat{h}^k(\tau_l^k) - h^k\|), \quad \text{as } l \to \infty. \tag{3.6.23}$$

This, together with the violation of (3.6.21), gives

$$(1 - \sigma_J)D_h\widehat{J}(h^k)(\widehat{h}^k(\tau_l^k) - h^k) + o(\|\widehat{h}^k(\tau_l^k) - h^k\|) > 0, \quad \text{as } l \to \infty. \tag{3.6.24}$$

Moreover, due to the relation (3.6.19), we also have

$$D_h\widehat{J}(h^k)(h^k - \widehat{h}^k(\tau_l^k)) \geq \frac{\|\widehat{h}^k(\tau_l^k) - h^k\|^2}{\tau_l^k}, \tag{3.6.25}$$

which further implies

$$o(\|\widehat{h}^k(\tau_l^k) - h^k\|) > (1 - \sigma_J)D_h\widehat{J}(h^k)(h^k - \widehat{h}^k(\tau_l^k)) \geq (1 - \sigma_J)\frac{\|\widehat{h}^k(\tau_l^k) - h^k\|^2}{\tau_l^k}, \quad \text{as } l \to \infty. \tag{3.6.26}$$

Thus, it follows from Lemma 3.6.6 that

$$\frac{\|\widehat{h}^k(\bar{\tau}) - h^k\|}{\bar{\tau}} \leq \frac{\|\widehat{h}^k(\tau_l^k) - h^k\|}{\tau_l^k} \to 0, \quad \text{as } l \to \infty. \tag{3.6.27}$$

This contradicts that h^k is not stationary. □

For the sake of our convergence analysis, we consider $\mathrm{tol}_h = \mathrm{tol}_\gamma = 0$ in the remainder of this section.

Lemma 3.6.8. *Let the sequence $\{(u^k, h^k, p^k) : k \in \mathcal{I}\}$ be generated by Algorithm 3.6.5. If $|\mathcal{I}|$ is infinite, then we have*

$$\liminf_{k\to\infty,\ k\in\mathcal{I}} \left\| h^k - \mathrm{Proj}_{Q_h}[h^k - \bar{\tau}D_h\widehat{J}(h^k)^\top] \right\| = 0. \tag{3.6.28}$$

Proof. We restrict ourselves to $k \in \mathcal{I}$ throughout this proof. It follows from Lemma 3.6.6 and the satisfaction of the Armijo-type condition (3.6.21) that

$$\begin{aligned}\widehat{J}(h^k) - \widehat{J}(h^{k+1}) &= \widehat{J}(h^k) - \widehat{J}(\widehat{h}^k(\tau^k)) \geq \sigma_J D_h \widehat{J}(h^k)(h^k - \widehat{h}^k(\tau^k)) \\ &\geq \sigma_J \frac{\|h^k - \widehat{h}^k(\tau^k)\|^2}{\tau^k} \geq \sigma_J \tau^k \frac{\|h^k - \widehat{h}^k(\bar{\tau})\|^2}{\bar{\tau}^2} \geq \frac{\sigma_J \underline{\tau}}{\bar{\tau}^2}\|h^k - \widehat{h}^k(\bar{\tau})\|^2,\end{aligned}$$

for all sufficiently large k. Moreover, since the sequence $\{\widehat{J}(h^k) : k \in \mathcal{I}\}$ is monotonically decreasing and $\widehat{J}$ is bounded from below, the conclusion follows. □

Now we are in a position to present the main result of our convergence analysis.

Theorem 3.6.9. *Let the sequence $\{(u^k, h^k)\}$ be generated by Algorithm 3.6.5. In addition, assume that the auxiliary variables $\{w^k\}$, recall (3.6.6) and (3.6.16) for the respective cases $k \in \mathcal{I}$ and $k \notin \mathcal{I}$ and also see equations (3.6.30) and (3.6.32) below, are uniformly bounded. Then there exists an accumulation point $\{(u^*, h^*)\}$ which is feasible and C-stationary for (3.3.7), i.e. $\{(u^*, h^*)\}$ satisfies (3.4.2) and (3.5.3) for some $p^* \in (\mathbb{R}^{|\Omega_u|})^2$, $w^* \in (\mathbb{R}^{|\Omega_u|})^2$, $v^* \in \mathbb{R}^{|\Omega_u|}$.*

Proof. The proof is divided into two cases.

Case I: Let us consider the case where $|\mathcal{I}|$ is infinite. In view of Lemma 3.6.8, let $\{(u^k, h^k, p^k)\}$ be a subsequence (the index k is kept throughout this proof for brevity) such that $k \in \mathcal{I}$ for all k and

$$\lim_{k\to\infty} \left\| h^k - \mathrm{Proj}_{Q_h}[h^k - \bar{\tau} D_h \widehat{J}(h^k)^\top] \right\| = 0. \tag{3.6.29}$$

Let (u^*, h^*, p^*) be an accumulation point of $\{(u^k, h^k, p^k)\}$. Note that (u^*, h^*, p^*) is feasible, i.e. satisfies the state equation (3.4.2), owing to the continuity of F and the closedness of G. If the strict complementarity holds at (u^*, h^*, p^*), then $\widehat{J}$ is continuously differentiable at h^*, and therefore we have $h^* = \mathrm{Proj}_{Q_h}[h^* - \bar{\tau} D_h \widehat{J}(h^*)^\top]$, or equivalently (u^*, h^*, p^*) is (C-)stationary.

Now assume that (u^*, h^*, p^*) lacks strict complementarity. For each k, let $g^k := D_h \widehat{J}(h^k)^\top$. Then from (3.6.5) we have

$$\left\{\begin{array}{l} g^k = D_h J(u^k, h^k)^\top + D_h F(u^k, h^k)^\top v^k, \\ D_u F(u^k, h^k)^\top v^k + \alpha \nabla^\top w^k + D_u J(u^k, h^k)^\top = 0, \\ w^k = \frac{1}{\alpha}(-I + p^k (p^k)^\top) \eta^k, \\ \alpha \nabla v^k + |\nabla u^k| \eta^k = 0, \end{array}\right. \tag{3.6.30}$$

with $v^k \to v^*$, $g^k \to g^*$, $w^k \to w^*$ as $k \to \infty$, possibly along yet another subsequence.

We claim that $(u^*, h^*, p^*, w^*, v^*)$ satisfies the C-stationarity (3.5.4). From (3.6.29) and (3.6.30), one readily verifies the first and the second conditions in (3.5.4). In view of the satisfaction of strict complementarity at each (u^k, h^k, p^k), the proof of the third condition that $(\alpha \nabla^\top w^*, -v^*) \in N^{(\mathrm{C})}_{\mathrm{gph}\, G}(u^*, \alpha \nabla^\top p^*)$ separates into two cases for each $j \in \Omega_u$.

(I-1) There exists a subsequence $\{(u^k, h^k, p^k)\}$ such that $(\nabla u^k)_j \neq 0$ and $|p^k_j| = 1$ for all k. Then it follows from (3.6.30) that

$$|(\nabla u^k)_j| w^k_j = (\nabla v^k)_j - \langle (\nabla v^k)_j, p^k_j \rangle p^k_j.$$

Analogous to (3.6.17), this eventually yields $\langle w_j^*, (\nabla v^*)_j\rangle \geq 0$ and $(\nabla v^*)_j - \langle(\nabla v^*)_j, p_j^*\rangle p_j^* = 0$.

(I-2) There exists a subsequence $\{(u^k, h^k, p^k)\}$ such that $(\nabla u^k)_j = 0$ and $|p_j^k| < 1$ for all k. Then it follows from (3.6.30) that $(\nabla v^*)_j = 0$.

In both cases above, $(\alpha\nabla^\top w^*, -v^*) \in N^{(\mathrm{C})}_{\mathrm{gph}\, G}(u^*, \alpha\nabla^\top p^*)$ holds true.

Case II: Now we turn to the case where $|\mathcal{I}|$ is finite. We claim that $\lim_{k\to\infty}\gamma^k = 0$ in this scenario. Assume for the sake of contradiction that for all sufficiently large k we have $\gamma^k = \bar\gamma$ for some $\bar\gamma > 0$ and $\|\breve{h}^k(\bar\tau) - h^k\| > \sigma_h\bar\gamma$. Then Algorithm 3.6.5, for all sufficiently large k, reduces to a projected gradient method on the constrained minimization (3.6.13) with a continuously differentiable objective. This leads to a contradiction as $\lim_{k\to\infty}\|\breve{h}^k(\bar\tau) - h^k\| = 0$ due to Proposition 2 in [GB82]. Thus, we must have $\lim_{k\to\infty}\gamma^k = 0$.

As a consequence, steps 17–23 in Algorithm 3.6.5 yields the existence of a subsequence $\{(u^k, h^k)\}$ such that $k \notin \mathcal{I}$ for all k and

$$\|\breve{h}^k(\bar\tau) - h^k\| = \left\|h^k - \mathrm{Proj}_{Q_h}[h^k - \bar\tau D_h\breve{J}_{\gamma^k}(h^k)^\top]\right\| \leq \sigma_h\gamma^k \to 0, \tag{3.6.31}$$

as $k\to\infty$. Let $g_\gamma^k := D_h\breve{J}_{\gamma^k}(h^k)^\top$, and we have

$$\begin{cases} F(u^k, p^k) + \alpha\nabla^\top p_\gamma^k = 0, \\ p_\gamma^k := \varphi'_{\gamma^k}(\nabla u^k), \\ g_\gamma^k = D_hJ(u^k, h^k)^\top + D_hF(u^k, h^k)^\top v^k, \\ D_uF(u^k, h^k)^\top v^k + \alpha\nabla^\top w^k = -D_uJ(u^k, h^k)^\top, \\ w^k = \varphi''_{\gamma^k}(\nabla u^k)\nabla v^k, \end{cases} \tag{3.6.32}$$

for all k such that $h^k \to h^*$, $u^k \to u^*$, $p_\gamma^k \to p^*$, $v^k \to v^*$, $w^k \to w^*$, $g_\gamma^k \to g_\gamma^*$ as $k \to \infty$, possibly along another subsequence. Then from (3.6.31) and (3.6.32), the first and the second conditions in the C-stationarity condition (3.5.4) immediately follow. The satisfaction of the third condition in (3.5.4) can be verified using an argument analogous to that in the proof of cases (1)–(3) in Theorem 3.6.4. Thus, we conclude that (u^*, h^*) is C-stationary. □

3.7 Numerical experiments

In this section, we report our numerical experiments on the bilevel optimization framework for blind deconvolution problems. In order to achieve practical efficiency, in section 3.7.1 we will utilize a simplified version of Algorithm 3.6.5. In particular, the smoothed lower-level problem can be efficiently handled by a semismooth Newton solver, which is described in section 3.7.1. Numerical results on PSF calibration and multiframe blind deconvolution are given in sections 3.7.2 and 3.7.3, respectively.

3.7.1 Implementation issues

Here our concern is to implement a practically efficient version of the hybrid projected gradient method (i.e. Algorithm 3.6.5) developed in section 3.6.3. At each iteration of that algorithm, step 14 requires the numerical solution of the set-valued equation (3.4.2) for obtaining a feasible point. In this vein, first-order methods are typically used, see, e.g., [CP11] and its variants, but they only converge sublinearly. We note that the semismooth Newton method without any regularization is not directly applicable for solving (3.4.2) due to non-uniqueness in the (dual) variable p. As a remedy, a null-space regularization on the predual problem is introduced in [HK04]. A more computationally amenable Tikhonov regularization (on the dual problem), which is equivalent to Huber-type smoothing on the primal objective, is proposed in [HS06]. Following [HS06], the Euler-Lagrange equation (3.6.11) in the smoothing step (i.e. steps 16–26) of Algorithm 3.6.5 can be solved by a superlinearly convergent semismooth Newton method. To take advantage of this fact, we will simplify Algorithm 3.6.5 by implementing the smoothing step only in Algorithm 3.7.2. In the meantime, we first describe a semismooth Newton solver for the smoothed lower-level problem.

Semismooth Newton solver for the smoothed lower-level problem

We only present essentials of the semismooth Newton method as a subroutine in solving the bilevel problem and refer the interested reader to [HS06, HW13, HW14b] for further details. For the smoothed lower-level problem in (3.6.10), we fix $\gamma > 0$ and $h \in Q_h$. With the introduction a dual variable $p^\gamma \in \left(\mathbb{R}^{|\Omega_u|}\right)^2$, the Euler-Lagrange equation (3.6.11) associated with the smoothing parameter γ can be reformulated as follows:

$$\begin{cases} (-\mu\Delta + K(h)^\top K(h))u^\gamma + \alpha\nabla^\top p^\gamma = K(h)^\top z, \\ \max(|\nabla u^\gamma|, \gamma)p^\gamma = \Big(\dfrac{3}{2} - \dfrac{|\nabla u^\gamma|^2}{2\max(|\nabla u^\gamma|, \gamma)^2}\Big)\nabla u^\gamma. \end{cases}$$

To ease our presentation, we temporarily omit the superscript γ in u^γ and p^γ, and denote the iterates in the lower-level solver (i.e. inner loop) by (u^l, p^l). A generalized Newton step on the above Euler-Lagrange equation refers to the solution of the following linear system:

$$\begin{bmatrix} -\mu\Delta + K(h)^\top K(h) & \alpha\nabla^\top \\ -C^l\nabla & \operatorname{diag}(m^l e) \end{bmatrix} \begin{bmatrix} \delta u^l \\ \delta p^l \end{bmatrix} = \begin{bmatrix} -(-\mu\Delta + K(h)^\top K(h))u^l - \alpha\nabla^\top p^l + K(h)^\top z \\ -m^l p^l + \Big(\dfrac{3}{2} - \dfrac{|\nabla u^l|^2}{2(m^l)^2}\Big)\nabla u^l \end{bmatrix},$$

where

$$\begin{aligned} m^l &:= \max(|\nabla u^l|, \gamma), \\ (\chi^l)_j &:= \begin{cases} 1 & \text{if } |(\nabla u^l)_j| \geq \gamma \\ 0 & \text{if } |(\nabla u^l)_j| < \gamma \end{cases} \quad \forall j \in \Omega_u, \\ C^l &:= \chi^l \left(I - (m^l)^{-1} p^l (\nabla u^l)^\top\right) + (1 - \chi^l)\left(\frac{3}{2}I - \operatorname{diag}\left(\frac{|\nabla u^l|^2 e}{2\gamma^2}\right) - \frac{(\nabla u^l)(\nabla u^l)^\top}{\gamma^2}\right). \end{aligned}$$

After eliminating δp^l in the above Newton system, we arrive at

$$\left(-\mu\Delta + K(h)^\top K(h) + \alpha\nabla^\top (m^l)^{-1} C^l \nabla\right) \delta u^l = -r(u^l; h, \gamma),$$

recall (3.6.11) for the definition of the residual term $r(\cdot)$. In order to guarantee that δu^l be a descent direction for the lower-level minimization problem, we further introduce a modification on C^l, i.e. we replace C^l by

$$\begin{aligned}\widehat{C}^l :=& \chi^l \left(I - \frac{1}{2}(m^l)^{-1}\left(\widehat{p}^l(\nabla u^l)^\top + (\nabla u^l)(\widehat{p}^l)^\top\right)\right) \\ &+ (1-\chi^l)\left(\frac{3}{2}I - \mathrm{diag}\left(\frac{|\nabla u^l|^2 e}{2\gamma^2}\right) - \frac{(\nabla u^l)(\nabla u^l)^\top}{\gamma^2}\right),\end{aligned}$$

where $\widehat{p}^l$ is the projection of p^l onto Q_p, i.e. $\widehat{p}^l := \frac{p^l}{\max(|p^l|,1)}$. Thus, the final modified Newton system appears as

$$\left(-\mu\Delta + K(h)^\top K(h) + \alpha\nabla^\top (m^l)^{-1} \widehat{C}^l \nabla\right) \delta u^l = -r(u^l; h, \gamma). \tag{3.7.1}$$

Once δu^l is obtained, δp^l can be computed by

$$\delta p^l := -p^l + (m^l)^{-1}\Big(\frac{3}{2} - \frac{|\nabla u^l|^2}{2(m^l)^2}\Big)\nabla u^l + (m^l)^{-1}\widehat{C}^l\nabla\delta u^l. \tag{3.7.2}$$

The overall semismooth Newton solver for the smoothed lower-level problem is summarized in Algorithm 3.7.1 below. The superlinear convergence of this solver can be justified following the approach in [HS06, HW13].

Algorithm 3.7.1 (Semismooth Newton solver)**.**

Require: (ordered) inputs $\alpha > 0$, $0 \le \mu \ll \alpha$, $h \in Q_h$, $\gamma > 0$, $u^1 \in \mathbb{R}^{|\Omega_u|}$, $\mathrm{tol}_r > 0$. **Return:** $u^* \in \mathbb{R}^{|\Omega_u|}$.

```
Initialize $p^1 := \frac{\nabla u^1}{\max(|\nabla u^1|, \gamma)}$, $l := 1$.
loop
    Generate the Newton system in (3.7.1).
    if $\frac{\|r(u^l; h, \gamma)\|}{\max(\|r(u^1; h, \gamma)\|, 1)} \le \mathrm{tol}_r$ then
        return $u^* := u^l$ and terminate the algorithm.
    end if
    Solve (3.7.1) for $\delta u^l$, and compute $\delta p^l$ using formula (3.7.2).
    Determine the step size $a^l > 0$ via backtracking Armijo line search along $\delta u^l$.
    Generate the next iterates: $u^{l+1} := u^l + a^l \delta u^l$ and $p^{l+1} := p^l + a^l \delta p^l$.
    Set $l := l + 1$.
end loop
```

Simplified projected gradient method

Based on Algorithm 3.7.1, we present the simplified projected gradient method for the bilevel problem (3.3.7) in the following. We remark that while the proximity measure κ^k in step 3 is chosen in our algorithm as a signal for reducing γ^k, other choices may be considered as well.

Algorithm 3.7.2 (Simplified projected gradient method)**.**

```
Require: inputs $\alpha > 0$, $0 \leq \mu \ll \alpha$, $\mathrm{tol}_r > 0$, $0 < \sigma_J < 1$, $\sigma_h > 0$, $\bar{\tau} > 0$, $\mathrm{tol}_\gamma > 0$, $0 < \rho_\gamma < 1$,
    $0 < \rho_\tau < 1$.
 1: Initialize $h^1 \in Q_h$, $\gamma^1 > 0$, $u^0 \in \mathbb{R}^{|\Omega_u|}$, $k := 1$.
 2: loop
 3:   Apply Algorithm 3.7.1 with ordered inputs $\alpha$, $\mu$, $h^k$, $\gamma^k$, $u^{k-1}$, $\mathrm{tol}_r$, which returns $u^k$ as
      the solution of (3.6.11).
 4:   Solve the adjoint equation
          $$\left(D_u F(u^k, h^k)^\top + \alpha \nabla^\top \varphi''_{\gamma^k}(\nabla u^k) \nabla\right) v^k = -D_u J(u^k, h^k)^\top$$
      for $v^k$. Then compute the gradient $D_h \breve{J}_{\gamma^k}(h^k)^\top := D_h J(u^k, h^k)^\top + D_h F(u^k, h^k)^\top v^k$ and
      evaluate the proximity measure
          $$\kappa^k := \left\| \mathrm{Proj}_{Q_h}[h^k - \bar{\tau} D_h \breve{J}_{\gamma^k}(h^k)^\top] - h^k \right\|.$$
 5:   if $\kappa^k \leq \sigma_h \gamma^k$ then
 6:     if $\gamma^k = \mathrm{tol}_\gamma$ then
 7:       return $(u^k, h^k)$ as a C-stationary point of (3.3.7) and terminate the algorithm.
 8:     else
 9:       Set $\gamma^{k+1} := \max(\rho_\gamma \gamma^k, \mathrm{tol}_\gamma)$. Go to step 13.
10:     end if
11:   end if
12:   Set $h^{k+1} := \mathrm{Proj}_{Q_h}[h^k - \tau^k D_h \breve{J}_{\gamma^k}(h^k)^\top]$, where $\tau^k$ the largest element in $\{\bar{\tau}(\rho_\tau)^l : l =$
      $0, 1, 2, ...\}$ which fulfills the following Armijo-type condition:
      $$\breve{J}_{\gamma^k}\left(\mathrm{Proj}_{Q_h}[h^k - \tau^k D_h \breve{J}_{\gamma^k}(h^k)^\top]\right) \leq \breve{J}_{\gamma^k}(h^k) + \sigma_J D_h \breve{J}_{\gamma^k}(h^k)(\mathrm{Proj}_{Q_h}[h^k - \tau^k D_h \breve{J}_{\gamma^k}(h^k)^\top] - h^k).$$
13:   Set $k := k + 1$.
14: end loop
```

We further specify the parameter choices for Algorithm 3.7.2 in our numerical experiments. For an image of $n_x \times n_y$ pixels, we set the mesh size $\omega := \sqrt{1/(n_x n_y)}$ and discretize the spatial gradient by forward differences, i.e. for each $j = (j_x, j_y) \in \Omega_u$

$$(\nabla u)_{(j_x, j_y)} := \left(\frac{u_{(j_x+1, j_y)} - u_{(j_x, j_y)}}{\omega}, \frac{u_{(j_x, j_y+1)} - u_{(j_x, j_y)}}{\omega} \right),$$

with homogenous Dirichlet boundary condition. The following parameters are chosen throughout the experiments: $\alpha = 10^{-5}$, $\mu = 10^{-4}\alpha$, $\sigma_J = \sigma_h = 0.01$, $\rho_\gamma = \rho_\tau = 1/2$, $u^0 = z$, $\gamma^1 = 0.05/\omega$, $\mathrm{tol}_\gamma = 0.001/\omega$, $\mathrm{tol}_r = 10^{-7}$. The conjugate gradient method is utilized for solving the linear systems in step 3 of Algorithm 3.7.1 with residual tolerance 0.01 and in step 3 of Algorithm 3.7.2 with residual tolerance 10^{-9}, respectively. All experiments are performed under Matlab R2011b.

3.7.2 Calibration of point spread functions

We first test our method on a point spread function (PSF) calibration problem. Let h be a point spread function on a 2D index domain Ω_h, and $Q_h = \{h \in \mathbb{R}^{|\Omega_h|} : \sum_{j\in\Omega_h} h_j = 1,\ h_j \geq 0\ \forall j \in \Omega_h\}$. The blurring operator K is defined through a 2D convolution, i.e. $K(h)u = h * u$, with zero boundary condition. Given the true PSF $h_{(\mathrm{true})} \in Q_h$ and the source image $u_{(\mathrm{true})} \in \mathbb{R}^{|\Omega_u|}$, the observed image z is generated as $h_{(\mathrm{true})} * u_{(\mathrm{true})} +$ noise, where the noise is white Gaussian and of zero mean and standard deviation 0.02. In addition to the observation, we are supplied with a reference image $u_{(\mathrm{ref})}$, which is generated as the (non-blurred) source image corrupted by white Gaussian noise of zero mean and standard deviation 0.02. Our aim is to calibrate the underlying PSF using a blurred observation image and a noisy reference image.

In this problem, we utilize a tracking-type objective

$$J(u,h) = \frac{1}{2}\|u - u_{(\mathrm{ref})}\|^2 + \frac{\beta}{2}\|\nabla h\|^2$$

in the upper level, where a Tikhonov regularization on h is also included to stabilize the solution and the regularization parameter $\beta = 0.05$ is chosen. The relevant partial derivatives of J and F required for the implementation of Algorithm 3.7.2 are listed below

$$\begin{aligned} D_u J(u,h)^\top &= u - u_{(\mathrm{ref})}, \\ D_h J(u,h)^\top &= -\beta\Delta h, \\ D_u F(u,h)^\top &= (-\mu\Delta + K(h)^\top K(h)), && (3.7.3) \\ \langle D_h F(u,h)^\top v, \delta h\rangle &= \langle v, D_h F(u,h)\delta h\rangle \\ &= \langle v, \delta h(-\cdot) * (h * u - z)\rangle + \langle v, h(-\cdot) * (\delta h * u)\rangle. && (3.7.4) \end{aligned}$$

Here $h(-\cdot)$ is a PSF in Q_h defined by $(h(-\cdot))_j = h_{-j}$ for all $j \in \Omega_h$, and similar for $\delta h(-\cdot)$. The size of Ω_h is always chosen to be slightly larger than the support size of the true PSF. Note that for $D_h F(u,h)^\top$ only the matrix-vector product $D_h F(u,h)^\top v$ is needed in the numerical computation, which is given by (3.7.4) in a dual form. Concerning the initializations, we set the initial line search step size $\bar{\tau} = 2 \times 10^{-5}$ and the initial PSF h^1 to be the discrete Dirac delta function.

Our experiments are performed on three different pairs of images and PSFs, namely Gaussian blur on the "Satellite" image, motion blur on the "Cameraman" image, and out-of-focus blur on the “Grain” image. In Figure 3.1, the ground-truth images are displayed in (a)–(c), the underlying PSFs in (d)–(f), and the corresponding blurred observations in (g)–(i). The results of the bilevel-optimization calibration are shown in the last two rows: (j)–(l) for the estimated PSFs and (m)–(o) for the deblurred images from the lower-level problem. It is observed that the calibrations are reasonably good in all three cases in the sense that the calibrated PSFs resemble their true counterparts and yield the deblurred images of high visual quality.

In Figure 3.2, we also illustrate the typical numerical behavior of Algorithm 3.7.2 in the “satellite” example. Subplot (a) records the history of the smoothing parameter γ^k. The objective values $J_{\gamma^k}(u^k, h^k)$ are shown in (b), which exhibit regular decrease along iterations. The proximity measure κ^k in step 4 of Algorithm 3.7.2, shown in subplot (c), also behaves well.

3.7.3 Multiframe blind deconvolution

Now we apply our algorithmic framework to *multiframe blind deconvolution* [CE07]. In this problem, the observation $\vec{z}$ consists of f frames, i.e. $\vec{z} = (\vec{z}^1, ..., \vec{z}^f)$, where each frame is generated from the convolution between the source image $u_{(\text{true})}$ and a frame-varying PSF $\vec{h}^i$ over Ω_h plus some additive Gaussian noise $\vec{\eta}^i$, i.e.

$$\vec{z}^i = \vec{h}^i * u_{(\text{true})} + \vec{\eta}^i, \quad \forall i \in \{1, 2, ..., f\}.$$

Furthermore, each PSF $\vec{h}^i$ follows a (normalized) multivariate Gaussian distribution, i.e. $\vec{h}^i = h(\vec{\sigma}_x^i, \vec{\sigma}_y^i, \vec{\theta}^i)$ with unknown frame-dependent parameters $\vec{\sigma}_x^i, \vec{\sigma}_y^i \in Q_\sigma$, $\vec{\theta}^i \in Q_\theta$. The parameterization of the Gaussian PSF $h : Q_\sigma \times Q_\sigma \times Q_\theta \to Q_h$ is defined by

$$h(\sigma_x, \sigma_y, \theta) := \frac{\widetilde{h}(\sigma_x, \sigma_y, \theta)}{\sum_{(j_x, j_y)\in\Omega_h} \left(\widetilde{h}(\sigma_x, \sigma_y, \theta)\right)_{(j_x, j_y)}},$$

where for all $(j_x, j_y) \in \Omega_h$

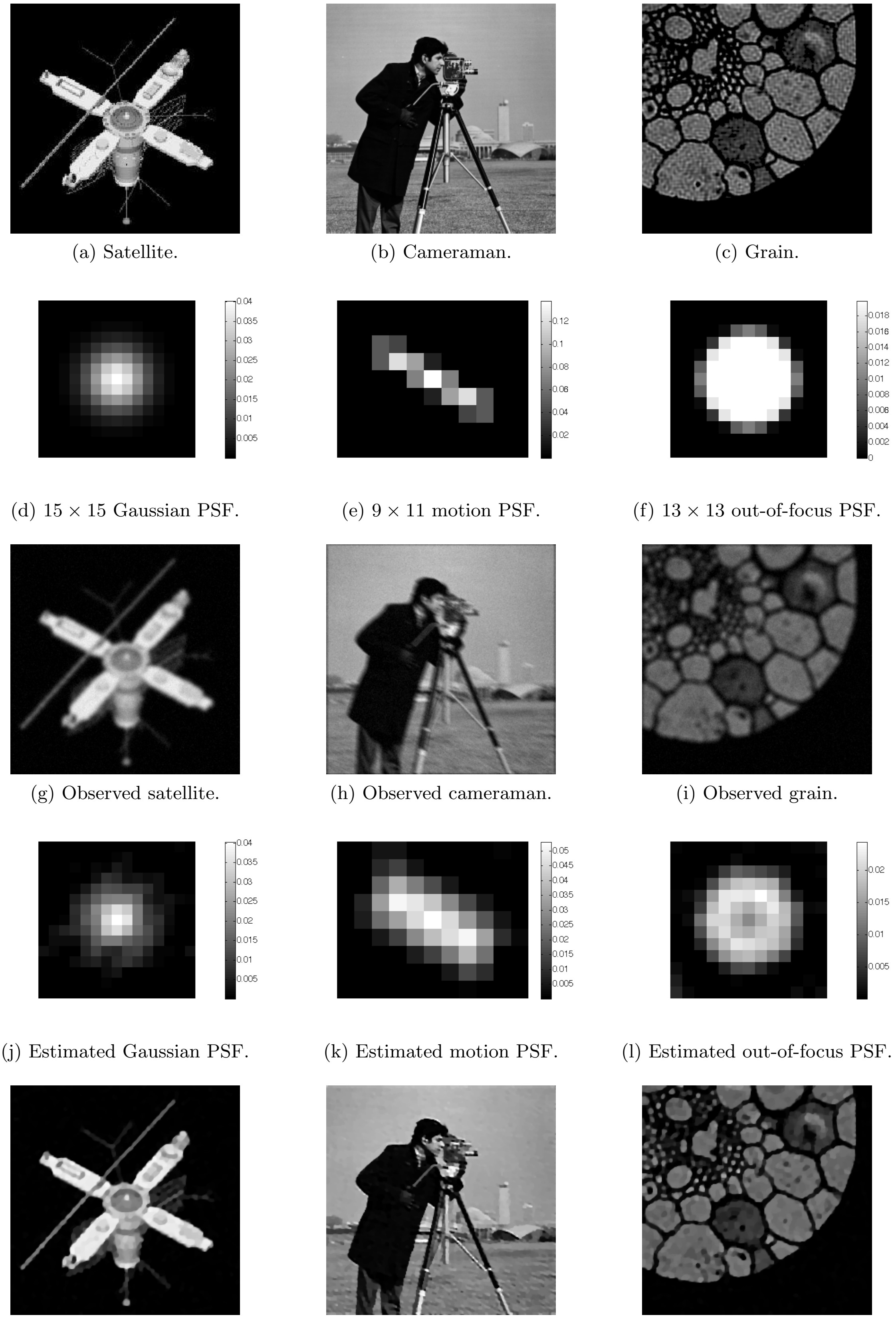
$$\left(\widetilde{h}(\sigma_x, \sigma_y, \theta)\right)_{(j_x, j_y)} := \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(j_x \cos\theta - j_y \sin\theta)^2}{2(\sigma_x)^2} - \frac{(j_x \sin\theta + j_y \cos\theta)^2}{2(\sigma_y)^2}\right).$$

Our task is to simultaneously recover the image $u_{(\text{true})}$ and the PSF parameters $\vec{\sigma}_x, \vec{\sigma}_y \in (Q_\sigma)^f$ and $\vec{\theta} \in (Q_\theta)^f$.

For such a multiframe blind deconvolution problem, we formulate the bilevel optimization model as follows:

$$\begin{aligned}
\min \quad & J(\vec{u}) = \frac{1}{2} \textstyle\sum_{k=1}^f \left\| \vec{u}^k - \frac{1}{f} \sum_{l=1}^f \vec{u}^l \right\|^2 \\
\text{s.t.} \quad & \vec{u}^i = \arg\min_{u\in\mathbb{R}^{|\Omega_u|}} \frac{1}{2} \left\| h(\vec{\sigma}_x^i, \vec{\sigma}_y^i, \vec{\theta}^i) * u - \vec{z}^i \right\|^2 + \alpha\|\nabla u\|_1, \quad \forall i \in \{1, 2, ...f\}, \\
& \vec{\sigma}_x, \vec{\sigma}_y \in (Q_\sigma)^f, \ \vec{\theta} \in (Q_\theta)^f.
\end{aligned}$$

The upper-level objective represents a (rescaled) sample variance of $\{\vec{u}^1, ..., \vec{u}^f\}$. Upon Huber-type smoothing on each lower-level problem respectively, the derivative of the reduced objective



(a) Satellite.

(b) Cameraman.

(c) Grain.

(d) 15 × 15 Gaussian PSF.

(e) 9 × 11 motion PSF.

(f) 13 × 13 out-of-focus PSF.

(g) Observed satellite.

(h) Observed cameraman.

(i) Observed grain.

(j) Estimated Gaussian PSF.

(k) Estimated motion PSF.

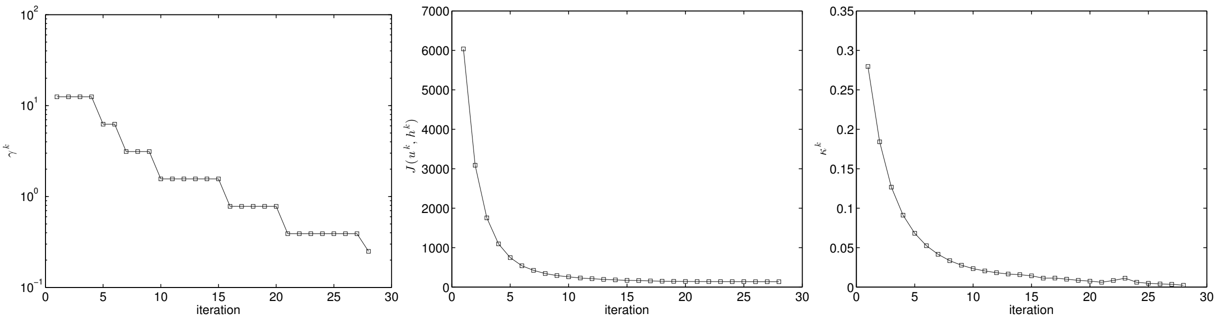
(l) Estimated out-of-focus PSF.

(m) Deblurred satellite.

(n) Deblurred cameraman.

(o) Deblurred grain.

Figure 3.1: Calibration of point spread functions.



(a) Smoothing parameter. (b) Objective value. (c) Proximity measure.

Figure 3.2: Numerical behavior.

$\widehat{J}(\vec{\sigma}_x, \vec{\sigma}_y, \vec{\theta}) := J(\vec{u}^1(\vec{\sigma}_x^1, \vec{\sigma}_y^1, \vec{\theta}^1), ..., \vec{u}^f(\vec{\sigma}_x^f, \vec{\sigma}_y^f, \vec{\theta}^f))$ can be calculated for all $i \in \{1, ..., f\}$ as

$$D_{(\vec{\sigma}_x^i, \vec{\sigma}_y^i, \vec{\theta}^i)} \widehat{J}(\vec{\sigma}_x, \vec{\sigma}_y, \vec{\theta})^\top = D_{(\sigma_x, \sigma_y, \theta)} h(\vec{\sigma}_x^i, \vec{\sigma}_y^i, \vec{\theta}^i)^\top D_h F(\vec{u}^i, \vec{h}^i)^\top \vec{v}^i,$$

where each $\vec{v}^i \in \mathbb{R}^{|\Omega_u|}$ satisfies the adjoint equation

$$\left(D_u F(\vec{u}^i, \vec{h}^i)^\top + \alpha \nabla^\top \varphi_\gamma''(\nabla \vec{u}^i) \nabla \right) \vec{v}^i = -D_{\vec{u}^i} J(\vec{u})^\top = -\left(\vec{u}^i - \frac{1}{f} \sum_{l=1}^{f} \vec{u}^l \right).$$

In addition, the formulae for $D_u F(\cdot)^\top$ and $D_h F(\cdot)^\top$ are identical to (3.7.3) and (3.7.4), and the partial derivatives of h are respectively given by

$$\begin{aligned}
\left(D_{\sigma_x} \widetilde{h}(\sigma_x, \sigma_y, \theta)^\top \right)_{(j_x, j_y)} &= \left(\widetilde{h}(\sigma_x, \sigma_y, \theta) \right)_{(j_x, j_y)} \left(\frac{(j_x \cos\theta - j_y \sin\theta)^2}{(\sigma_x)^3} - \frac{1}{\sigma_x} \right), \\
D_{\sigma_x} h(\sigma_x, \sigma_y, \theta)^\top &= \frac{1}{\sum_{(j_x, j_y) \in \Omega_h} \left(\widetilde{h}(\sigma_x, \sigma_y, \theta) \right)_{(j_x, j_y)}} \cdot \\
&\quad \left(D_{\sigma_x} \widetilde{h}(\sigma_x, \sigma_y, \theta)^\top - h(\sigma_x, \sigma_y, \theta) \sum_{(j_x, j_y) \in \Omega_h} \left(D_{\sigma_x} \widetilde{h}(\sigma_x, \sigma_y, \theta)^\top \right)_{(j_x, j_y)} \right), \\
\left(D_{\sigma_y} \widetilde{h}(\sigma_x, \sigma_y, \theta)^\top \right)_{(j_x, j_y)} &= \left(\widetilde{h}(\sigma_x, \sigma_y, \theta) \right)_{(j_x, j_y)} \left(\frac{(j_x \sin\theta + j_y \cos\theta)^2}{(\sigma_y)^3} - \frac{1}{\sigma_y} \right), \\
D_{\sigma_y} h(\sigma_x, \sigma_y, \theta)^\top &= \frac{1}{\sum_{(j_x, j_y) \in \Omega_h} \left(\widetilde{h}(\sigma_x, \sigma_y, \theta) \right)_{(j_x, j_y)}} \cdot \\
&\quad \left(D_{\sigma_y} \widetilde{h}(\sigma_x, \sigma_y, \theta)^\top - h(\sigma_x, \sigma_y, \theta) \sum_{(j_x, j_y) \in \Omega_h} \left(D_{\sigma_y} \widetilde{h}(\sigma_x, \sigma_y, \theta)^\top \right)_{(j_x, j_y)} \right), \\
\left(D_\theta \widetilde{h}(\sigma_x, \sigma_y, \theta)^\top \right)_{(j_x, j_y)} &= \left(\widetilde{h}(\sigma_x, \sigma_y, \theta) \right)_{(j_x, j_y)} \left(\frac{1}{(\sigma_x)^2} - \frac{1}{(\sigma_y)^2} \right) \cdot \\
&\quad (j_x \cos\theta - j_y \sin\theta)(j_x \sin\theta + j_y \cos\theta),
\end{aligned}$$

$$D_\theta h(\sigma_x,\sigma_y,\theta)^\top = \frac{1}{\sum_{(j_x,j_y)\in\Omega_h}\left(\widetilde{h}(\sigma_x,\sigma_y,\theta)\right)_{(j_x,j_y)}}\cdot$$
$$\left(D_\theta\widetilde{h}(\sigma_x,\sigma_y,\theta)^\top - h(\sigma_x,\sigma_y,\theta)\sum_{(j_x,j_y)\in\Omega_h}\left(D_\theta\widetilde{h}(\sigma_x,\sigma_y,\theta)^\top\right)_{(j_x,j_y)}\right).$$

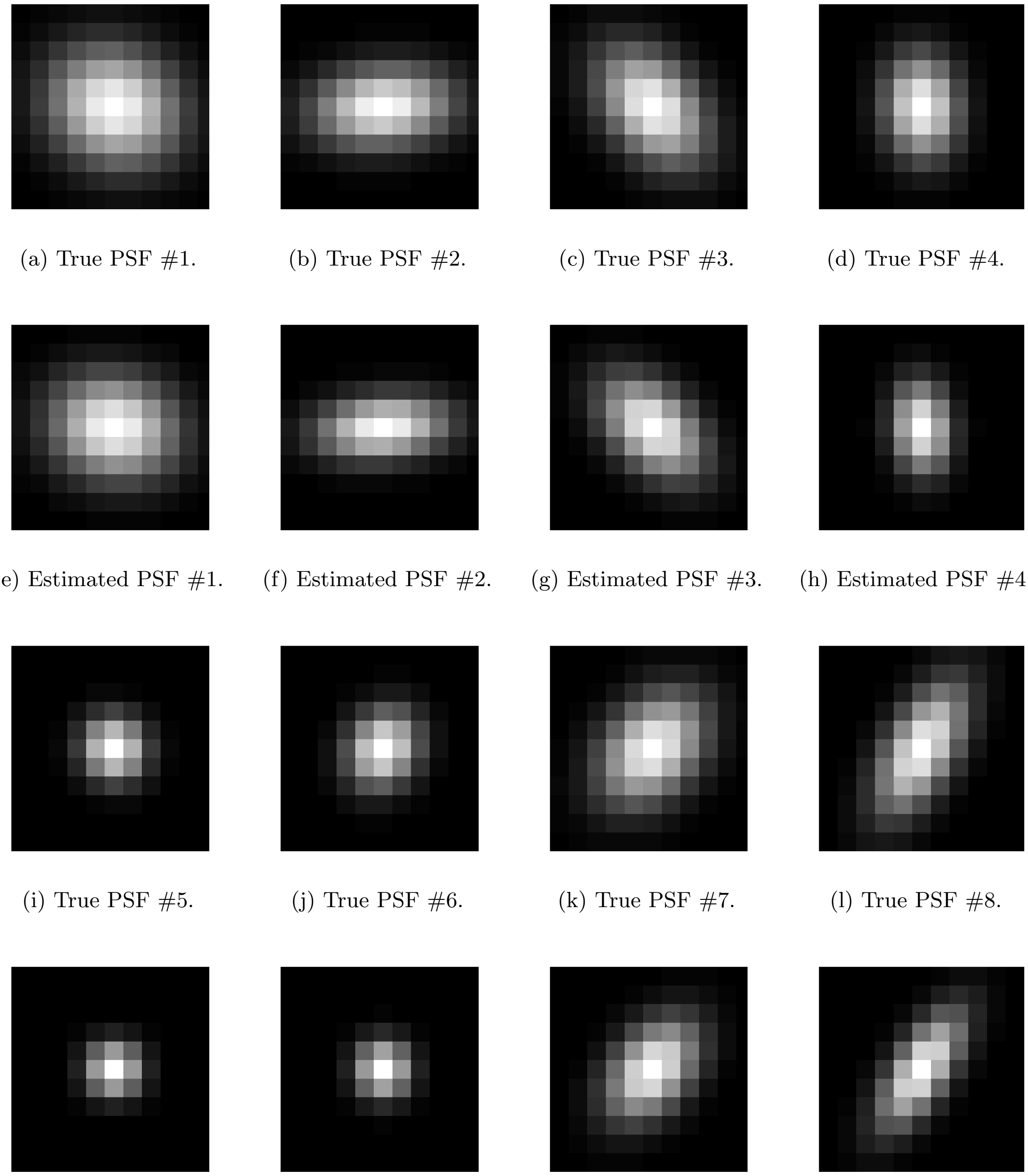
In our experiments, $Q_\sigma = [1,3]$ and $Q_\theta = [-\pi/2, \pi/2]$ are fixed, and the underlying parameters $(\vec{\sigma}_x^{(\text{true})}, \vec{\sigma}_y^{(\text{true})}, \vec{\theta}^{(\text{true})})$ are (uniform-)randomly drawn from $(Q_\sigma)^f \times (Q_\sigma)^f \times (Q_\theta)^f$. The first and third rows of Figure 3.3 show the random PSFs in a trial run with 8 frames, i.e. $f = 8$. The corresponding observations are given in the first and third rows of Figure 3.4. Concerning the initializations in our implementation, we always choose $\bar{\tau} = 0.005$ and $(\vec{\sigma}_x^i)^1 = (\vec{\sigma}_y^i)^1 = 2$, $(\vec{\theta}^i)^1 = 0$ for all i.

The results of the 8-frame trial run, both PSFs and images, are displayed in Figures 3.3 and 3.4 respectively. It is observed from the comparison in Figure 3.3 that our method well captures the underlying PSFs, especially the widths and the orientations in case of strongly skewed PSFs (see #2, #3, #4, #7, #8). Furthermore, all deblurred frames yield significant improvement in visual quality over the corresponding observations.

We are also interested in the effect of the number of frames on the image restoration quality. For this sake, we track the mean peak signal-to-noise ratio (mPSNR) of all individual frames for $f \in \{4, 6, 8, 10, 12\}$. For each f, the mean and the standard deviation (stdev) of mPSNR after 10 trial runs are reported in Table 3.1, where the mean is rising and the standard deviation is falling as f becomes larger. Thus, we conclude from our experiments that, as is expected, more observations typically enhance the frame-wise image restoration quality in the bilevel-optimization based multiframe blind deconvolution.

f	4	6	8	10	12
mean	23.6019	23.7170	23.7639	23.7883	24.0026
stdev	0.6020	0.4380	0.3381	0.2889	0.2720

Table 3.1: Mean peak signal-to-noise ratio.



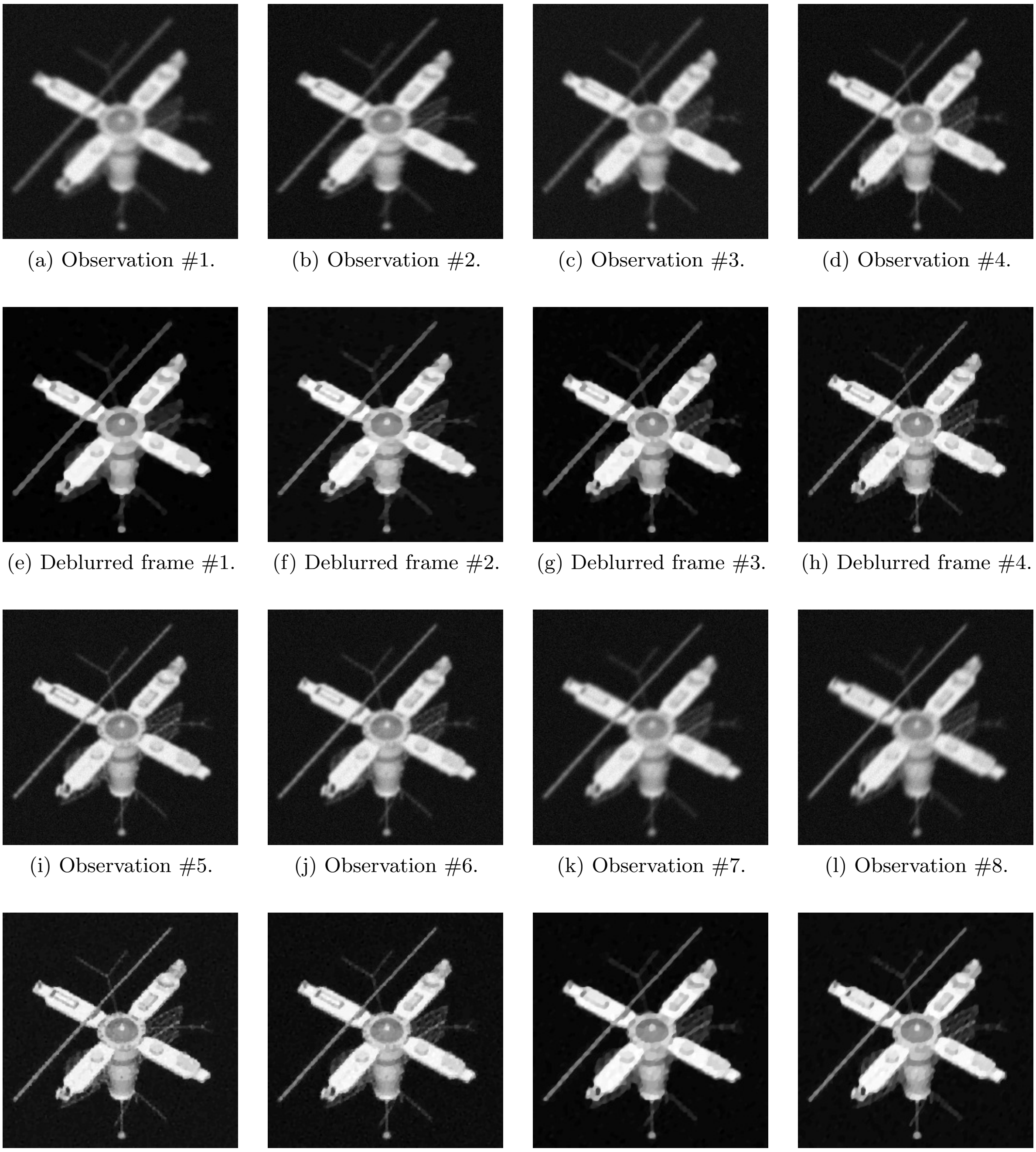
(a) True PSF #1. (b) True PSF #2. (c) True PSF #3. (d) True PSF #4.

(e) Estimated PSF #1. (f) Estimated PSF #2. (g) Estimated PSF #3. (h) Estimated PSF #4.

(i) True PSF #5. (j) True PSF #6. (k) True PSF #7. (l) True PSF #8.

(m) Estimated PSF #5. (n) Estimated PSF #6. (o) Estimated PSF #7. (p) Estimated PSF #8.

Figure 3.3: Multiframe blind deconvolution — PSFs.



(a) Observation #1. (b) Observation #2. (c) Observation #3. (d) Observation #4.

(e) Deblurred frame #1. (f) Deblurred frame #2. (g) Deblurred frame #3. (h) Deblurred frame #4.

(i) Observation #5. (j) Observation #6. (k) Observation #7. (l) Observation #8.

(m) Deblurred frame #5. (n) Deblurred frame #6. (o) Deblurred frame #7. (p) Deblurred frame #8.

Figure 3.4: Multiframe blind deconvolution — images.

Chapter 4

Robust principal component pursuit: a Riemannian optimization approach

This chapter is concerned with low-rank matrices, which can be viewed as being sparse in singular values. Such a distinct nature of sparsity leads to very different variational models as well as numerical techniques in comparison with the previous two chapters.

4.1 Introduction

A typical approach in understanding big and complex data in many different application areas utilizes data decomposition additively splitting the given data into several components of respective low complexity. For this purpose, robust principal component pursuit (RPCP), introduced in [CLMW11], aims at recovering a low-rank component and a sparse component from a possibly noisy data matrix. The low-rank component often refers to a certain intrinsically low dimensional pattern in the data, while the sparse component corresponds to either grossly corrupted measurements or pattern-irrelevant data. In this sense, RPCP is more robust in practice than the classical principal component analysis. The RPCP and its variants have found various promising applications, particularly in image and signal processing; e.g. video surveillance [CLMW11], face recognition [JCM12], texture modeling [ZGLM12], video inpainting [JHSX11], audio separation [HCSHJ12], latent semantic indexing [MZWM10], etc.

Concerning the numerical solution of RPCP in the large-scale setting, a popular approach [CLMW11, CSPW11] is to solve a "relaxed" convex program, where the rank functional is relaxed by the nuclear norm, i.e. the sum of the singular values, and the cardinality function is relaxed by the ℓ^1-norm, i.e. the sum of all entries in absolute values. In [CLMW11], it was proven that the convex-relaxation model provides the exact recovery with dominating probability given some mild assumptions on the underlying low-rank and sparse components. A somewhat more deterministic argument can be found in [CSPW11], where a sufficient condition for exact recovery, based on the notion of rank-sparsity incoherence, is invoked. This condition holds true with high probability for random low-rank and sparse components. Based on the convex-relaxation formulation, for the numerical solution of the associated minimization problem an augmented Lagrangian method (ALM) is utilized in [CLMW11]. A related work on ALM, improving efficiency of the method and expanding its scope with respect to applications, can be found in [TY11]. A list of works concerning numerical solvers relevant to the convex-relaxation approach is contained in [LRM]. Typically, at each iteration such solvers involve the computation of a singular value decomposition (SVD) in full dimension, which becomes highly expensive in large-scale applications. Acceleration of this SVD step can be possibly done via a Lancoz-based partial SVD technique (see, e.g., [PRO] for an efficient implementation under Matlab) as suggested in [CCS10, TY11], but its practical efficiency largely relies on the properties of the target matrix of the SVD such as relatively low rank and/or fast matrix-vector multiplication. Finally, besides the convex-relaxtion based approaches, we also mention a (nonconvex) factorization-based augmented Lagrangian alternating direction method for RPCP [WYZ12], for which an online code is available [LMa].

In this work, we solve RPCP by formulating a (regularized) least-squares problem with rank and cardinality constraints; see (4.3.1) below. Then an alternating minimization scheme (AMS) is employed to seek a stationary point which satisfies the first-order necessary optimality condition. If each subproblem in AMS is solved exactly by global minimization (i.e. metric projection) along the iterations, then AMS essentially becomes a heuristic method of (generalized) alternating projection onto manifolds (see the Appendix and also [LM08]), which is known to be locally convergent for transversal manifolds. However, the convergence of this alternating projection method can be possibly spoiled by defective initial guesses, which calls upon proper globalization (or safeguard) strategies on AMS.

For this sake, we propose a general framework sufficient for AMS to converge globally, which is then activated algorithmically. In particular, the low-rank subproblem is solved inexactly by a Riemannian (manifold) optimization step such that SVDs in full dimension can be favorably avoided. We point out that Riemannian optimization is an active research area in its own right; see [AMS08] and the references therein for an introduction on the subject and [BMAS14] for a miscellaneous toolbox available online. Concerning the applications of Riemannian optimization related to low-rank matrices, we refer to [KMO10a, KMO10b, SE10, BA11, MBS11, Van13] among other references which appeared very recently. Nevertheless, most of these papers, if not all, address the context of *low-rank matrix completion* [CR09] rather than RPCP, i.e. the sparse component is of no concern. In the present work, however, we include such a sparse component (in addition to the low-rank part) by embedding a tailored Riemannian optimization technique, namely the projected dogleg step, into the overall AMS. A q-linear convergence theory is established from the perspective of an inexact Newton method on the underlying matrix manifold. For the implementation of AMS, we also propose a heuristic trimming procedure which performs a proper tuning of the underlying rank and cardinality constraints. This procedure aims at automatically identifying the appropriate rank and sparsity of the two target components within the given data.

The remainder of the chapter is organized as follows. Preliminaries on Riemannian optimization, as a major component of our algorithmic development later on, are provided in section 4.2. In section 4.3, we formulate our variational model for RPCP and investigate the existence of a solution as well as the first-order optimality condition. The overall AMS and its convergence analysis are presented in detail in section 4.4. Section 4.5 concludes the chapter with a series of numerical experiments on the proposed method, including a comparison with a currently state-of-the-art augmented Lagrangian method. An appendix on local convergence of an alternating projection method for RPCP is attached in section 4.6.

4.2 Preliminaries on Riemannian optimization

In this section, we provide a concise review on the essential elements of differential geometry in a general context, which facilitate the Riemannian optimization technique used in our algorithmic development later on. Most of the presented materials can be found in standard differential geometry textbooks [Boo03, O'n83]. The concepts related to Riemannian optimization, e.g. Riemannian gradient, Riemannian Hessian, and retraction, are less standard in the literature, on which we refer to the monograph [AMS08] for a more comprehensive introduction.

Smooth manifold

Let $\mathcal{M}$ be an n-dimensional smooth manifold. Given any $p \in \mathcal{M}$, there exists a homeomorphism φ (termed *a local chart*) mapping from a neighborhood U of p on $\mathcal{M}$ onto an open subset $\varphi(U)$ in $\mathbb{R}^n$. Sometimes it is also convenient to denote $\varphi = (x^1, ..., x^n)$ where each $x^i : U \to \mathbb{R}$ is a coordinate function. If $\mathcal{M} = \mathbb{R}^n$, for any $p = (p_1, ..., p_n)$, we call $u^i : p \in \mathbb{R}^n \mapsto p_i \in \mathbb{R}$ a natural coordinate function. Thus, each coordinate function can be understood as a composition $x^i = u^i \circ \varphi$.

Tangent vector

In differential geometry, a tangent vector v is often considered as a derivation which maps from a smooth function (or a scalar field) f on the manifold $\mathcal{M}$ to the directional derivative of f along v. Denote the set of all scalar fields on $\mathcal{M}$ by $\mathfrak{F}(\mathcal{M})$. Then a tangent vector of $\mathcal{M}$ at p is formally defined as a mapping $v : \mathfrak{F}(\mathcal{M}) \to \mathbb{R}$ such that

1. v is $\mathbb{R}$-linear, i.e. $v(af + bg) = av(f) + bv(g)$ for any $a, b \in \mathbb{R}$.

2. v is Leibnizian, i.e. $v(fg) = v(f)g(p) + v(g)f(p)$ for any $f, g \in \mathfrak{F}(\mathcal{M})$.

Note that all tangent vectors at p form a vector space. We call this vector space, denoted by $T_{\mathcal{M}}(p)$, the tangent space to $\mathcal{M}$ at p. Given a local chart φ around p, one can define a coordinate basis vector $\partial_i(p)$ for the tangent space $T_{\mathcal{M}}(p)$ by

$$\partial_i(p)f = \frac{\partial(f \circ \varphi^{-1})}{\partial u^i}(\varphi(p)) \quad \text{for all } f \in \mathfrak{F}(\mathcal{M}),$$

where u^i is a natural coordinate function. It turns out that $\{\partial_i(p) : i = 1, ..., n\}$ forms a basis for $T_{\mathcal{M}}(p)$, and we have for any $v \in T_{\mathcal{M}}(p)$ that $v = \sum_{i=1}^n v(x^i)\partial_i(p)$; see Theorem 1.12 in [O'n83].

In accordance with a tangent vector, a vector field ξ on $\mathcal{M}$ is a smooth function which maps any $p \in \mathcal{M}$ to a tangent vector $v \in T_{\mathcal{M}}(p)$. The set of all vector fields on $\mathcal{M}$ is denoted by $\mathfrak{X}(\mathcal{M})$.

Riemannian metric

The first-order geometry on a manifold requires the notion of *Riemannian metric*. Let each tangent space $T_{\mathcal{M}}(p)$ be endowed with an inner product $\langle \cdot, \cdot \rangle_p$, which is a bilinear, symmetric positive definite form. If $\langle \cdot, \cdot \rangle_p$ is smoothly varying with p over $\mathcal{M}$, then we call $\langle \cdot, \cdot \rangle$ a Riemannian metric and $(\mathcal{M}, \langle \cdot, \cdot \rangle)$ a Riemannian manifold. In fact, any second-countable Hausdorff manifold admits a Riemannian metric; see pp. 45 in [AMS08]. Given any two vector fields $\xi, \eta \in \mathfrak{X}(\mathcal{M})$ expanded in coordinate vector fields, i.e. $\xi = \sum_{i=1}^n \xi^i \partial_i$ and $\eta = \sum_{j=1}^n \eta^j \partial_j$, the Riemannian metric $\langle \cdot, \cdot \rangle$ can be encoded by $\langle \xi, \eta \rangle = \sum_{i,j} \langle \partial_i, \partial_j \rangle \xi^i \eta^j$ in matrix form with entries $\{\langle \partial_i, \partial_j \rangle : i, j = 1, ..., n\}$.

Riemannian gradient

On a Riemannian manifold $(\mathcal{M}, \langle \cdot, \cdot \rangle)$, the *Riemannian gradient* of a scalar field $f \in \mathfrak{F}(\mathcal{M})$, denoted by $\text{grad} f$, is defined as a vector field on $\mathcal{M}$ such that

$$\langle \text{grad} f, \xi \rangle = \xi f, \tag{4.2.1}$$

for all $\xi \in \mathfrak{X}(\mathcal{M})$. When $\mathcal{M}$ is an embedded submanifold, $\text{grad} f$ can be calculated through an orthogonal projection as in the following theorem; see pp. 48 in [AMS08].

Theorem 4.2.1. *Let $\mathcal{M}$ be an embedded submanifold of a Riemannian manifold $(\widehat{\mathcal{M}}, \langle \cdot, \cdot \rangle)$ endowed with the induced metric $\langle \cdot, \cdot \rangle$. Let f be the restriction of a scalar field $\widehat{f} \in \mathfrak{F}(\widehat{\mathcal{M}})$ on $\mathcal{M}$, i.e. $f(p) = \widehat{f}(p)$ for all $p \in \mathcal{M}$. Then we have for all $p \in \mathcal{M}$ that*

$$\text{grad} f(p) = \text{Proj}_{T_{\mathcal{M}}(p)}(\text{grad}\widehat{f}(p)),$$

where $\text{Proj}_{T_{\mathcal{M}}(p)}$ denotes the orthogonal projection from $T_{\widehat{\mathcal{M}}}(p)$ to $T_{\mathcal{M}}(p)$.

Proof. The conclusion follows since $\text{Proj}_{T_{\mathcal{M}}(p)}(\text{grad}\widehat{f}(p)) \in T_{\mathcal{M}}(p)$ and for all $v \in T_{\mathcal{M}}(p)$ we have $\langle \text{Proj}_{T_{\mathcal{M}}(p)}(\text{grad}\widehat{f}(p)), v \rangle = \langle \text{grad}\widehat{f}(p), v \rangle = (v\widehat{f})(p) = (vf)(p) = \langle \text{grad} f(p), v \rangle$. □

Riemannian connection

The second-order geometry on a manifold relies on the notion of the Riemannian connection. Any Riemannian manifold $(\mathcal{M}, \langle\cdot,\cdot\rangle)$ admits a unique *Riemannian connection* $\nabla : (\xi, \eta) \in \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \mapsto \nabla_\eta \xi \in \mathfrak{X}(\mathcal{M})$ such that the following properties are satisfied for any $\xi, \eta, \zeta \in \mathfrak{X}(\mathcal{M})$, $f, g \in \mathfrak{F}(\mathcal{M})$, and $a, b \in \mathbb{R}$:

1. ∇ is $\mathfrak{F}(\mathcal{M})$-linear in η, i.e. $\nabla_{f\eta+g\zeta}\xi = f\nabla_\eta\xi + g\nabla_\zeta\xi$.

2. ∇ is $\mathbb{R}$-linear in ξ, i.e. $\nabla_\eta(a\xi + b\zeta) = a\nabla_\eta\xi + b\nabla_\eta\zeta$.

3. ∇ is Leibnizian, i.e. $\nabla_\eta(f\xi) = (\eta f)\xi + f\nabla_\eta\xi$.

4. ∇ is symmetric, i.e. $(\nabla_\eta\xi)f - (\nabla_\xi\eta)f = \eta(\xi f) - \xi(\eta f)$.

5. ∇ is metric-compatible, i.e. $\zeta\langle\xi,\eta\rangle = \langle\nabla_\zeta\xi,\eta\rangle + \langle\xi,\nabla_\zeta\eta\rangle$.

See Theorem 3.11 in [O'n83]. The vector field $\nabla_\eta\xi$ is termed a *covariant derivative.*

It is often convenient to represent a covariant derivative in terms of Christoffel symbols. Given a local chart on a neighborhood U of $\mathcal{M}$ and $\xi = \sum_{i=1}^n \xi^i\partial_i$, $\eta = \sum_{j=1}^n \eta^j\partial_j$ expanded in coordinate vector fields, we write for each $p \in U$ that

$$\nabla_\eta\xi(p) = \sum_{i,j,k} \Gamma_p^{ijk}\eta^j(p)\xi^i(p)\partial_k(p) + \sum_{i,j} \eta^j(p)(\partial_j\xi^i)(p)\partial_i(p),$$

where the *Christoffel symbols* $\{\Gamma_p^{ijk} \in \mathbb{R} : i, j, k = 1, ..., n\}$ are defined by $\nabla_{\partial_j}\partial_i(p) = \sum_k \Gamma_p^{ijk}\partial_k(p)$ for all $p \in U$ and $i, j = 1, ..., n$. In the remainder of our presentation, we prefer a matrix notation of the Christoffel symbols (see, e.g., [EAS98]), i.e.

$$\nabla_\eta\xi(p) = D\xi(p)[\eta(p)] + \Gamma_p[\xi(p), \eta(p)] \tag{4.2.2}$$

holds for $p \in U$ and a symmetric, $\mathfrak{F}(\mathcal{M})$-bilinear map $\Gamma_p : T_\mathcal{M}(p) \times T_\mathcal{M}(p) \to T_\mathcal{M}(p)$. Here $D\xi(p)[\eta(p)]$ denotes the directional derivative of ξ along $\eta(p)$ at p.

Parallel translation, geodesic, and exponential mapping

Let $\gamma : t \in Q(\subset \mathbb{R}) \mapsto \gamma(t) \in \mathcal{M}$ be a smooth curve, parameterized over an interval Q, on the Riemannian manifold $(\mathcal{M}, \langle\cdot,\cdot\rangle)$ equipped with the Riemannian connection ∇. Denote the set of all scalar fields and the set of all vector fields on the curve γ by $\mathfrak{F}(\gamma)$ and $\mathfrak{X}(\gamma)$ respectively. In particular, the velocity field $\dot\gamma$ defined by $\dot\gamma : f \in \mathfrak{F}(\gamma) \mapsto \frac{d}{dt}(f \circ \gamma) \in \mathbb{R}$ is a vector field on γ. It can be shown, see Proposition 3.18 in [O'n83], that there is a unique function $\frac{D}{dt} : \mathfrak{X}(\gamma) \to \mathfrak{X}(\gamma)$ such that the following properties are satisfied for any $\xi, \eta \in \mathfrak{X}(\gamma)$, $\zeta \in \mathfrak{X}(\mathcal{M})$, $f \in \mathfrak{F}(\gamma)$, and $a, b \in \mathbb{R}$:

1. $\frac{D}{dt}(a\xi + b\eta) = a\frac{D}{dt}\xi + b\frac{D}{dt}\eta$.

2. $\frac{D}{dt}(f\xi) = \frac{df}{dt}\xi + f\frac{D}{dt}\xi$.

3. $\frac{D}{dt}(\zeta \circ \gamma)(\tau) = \nabla_{\dot{\gamma}(\tau)}\zeta(\gamma(\tau))$ for each $\tau \in Q$.

4. $\frac{d}{dt}\langle \xi, \eta \rangle = \langle \frac{D}{dt}\xi, \eta \rangle + \langle \xi, \frac{D}{dt}\eta \rangle$.

We call $\frac{D}{dt}\xi$ the *induced covariant derivative* of ξ on γ. Given a local chart around $\gamma(\tau)$, the induced covariant derivative can be represented by the Christoffel symbol as

$$\frac{D}{dt}\xi(\tau) = \frac{d}{dt}\xi(\tau) + \Gamma_{\gamma(\tau)}[\dot{\gamma}(\tau), \xi(\tau)].$$

The vector field $\xi \in \mathfrak{X}(\gamma)$ is said to be *parallel* along γ if $\frac{D}{dt}\xi = 0$ everywhere on Q. For a curve $\gamma : Q \to \mathcal{M}$ with $\tau_0 \in Q$ and $v_0 \in T_{\mathcal{M}}(\gamma(\tau_0))$, it follows from the fundamental theorem of existence and uniqueness for ordinary differential equations that there exists a unique parallel vector field $\pi \in \mathfrak{X}(\gamma)$ along γ such that $\pi(\tau_0) = v_0$; see Proposition 3.19 in [O'n83]. We call the operator $v_0 \in T_{\mathcal{M}}(\gamma(\tau_0)) \mapsto \pi(\tau_1) \in T_{\mathcal{M}}(\gamma(\tau_1))$ the *parallel translation* on $\mathcal{M}$ along γ from $\gamma(\tau_0)$ to $\gamma(\tau_1)$.

Geodesics generalize straight lines in Euclidean spaces. Define the acceleration field $\ddot{\gamma}$ on γ by $\ddot{\gamma} = \frac{D}{dt}\dot{\gamma}$. Then γ qualifies as a geodesic if $\ddot{\gamma}(\tau) = 0$ for every $\tau \in Q$. In terms of the Christoffel symbols, $\ddot{\gamma}(\tau) = 0$ is equivalent to

$$\frac{d^2}{dt^2}\gamma(\tau) + \Gamma_{\gamma(\tau)}[\dot{\gamma}(\tau), \dot{\gamma}(\tau)] = 0.$$

For $p = \gamma(0)$ and $v \in T_{\mathcal{M}}(p)$, there exists an interval Q containing 0 and a unique *geodesic* $\gamma : Q \to \mathcal{M}$ such that $\dot{\gamma}(0) = v$; see Lemma 3.22 in [O'n83]. We call such a curve $\gamma(t; p, v)$ a geodesic starting at p with initial velocity v. Note that $t \mapsto \gamma(t; p, v)$ is homogenous in the sense that, as long as $t, at \in Q$, we have $\gamma(at; p, v) = \gamma(t; p, av)$ for any $a \in \mathbb{R}$.

Thus far, we are able to define the *exponential mapping* at p by $\exp_p : v \in T_{\mathcal{M}}(p) \mapsto \gamma(1; p, v) \in \mathcal{M}$. It can be easily verified that the differential of the exponential mapping at $0_p \in T_{\mathcal{M}}(p)$ is an identity map on $T_{\mathcal{M}}(p)$, i.e. $D\exp_p(0_p)[v] = v$ for all $v \in T_{\mathcal{M}}(p)$. Consequently, $\exp_p$ is a local diffeomorphism from $T_{\mathcal{M}}(p)$ to $\mathcal{M}$ around 0_p; see Proposition 3.30 in [O'n83].

Riemannian Hessian

Riemannian Hessian is central to second-order methods in Riemannian optimization, e.g. Riemannian Newton method and Riemannian trust-region method; see [Smi93, EAS98, AMS08]. On a Riemannian manifold $(\mathcal{M}, \langle \cdot, \cdot \rangle)$ equipped with the Riemannian connection ∇, the *Riemannian Hessian* of $f \in \mathfrak{F}(\mathcal{M})$ at $p \in \mathcal{M}$, denoted by $\text{Hess}f(p)$, is a linear mapping from $T_{\mathcal{M}}(p)$ to itself such that

$$\text{Hess}f(p)[\xi(p)] = \nabla_{\xi}\text{grad}f(p), \tag{4.2.3}$$

for each $\xi \in \mathfrak{X}(\mathcal{M})$.

In comparison with the Riemannian gradient, the explicit calculation of Riemannian Hessian is in general more involved, for which an alternative is to make use of a retraction.

Retraction

Conceptually, a retraction is a map from the tangent space of the manifold to the manifold itself which, to some extent, behaves locally as the exponential mapping. In fact, the exponential mapping, as a special example of retractions, played a dominating role in the early literature of Riemannian optimization. However, very often the exponential mapping, which involves the solutions of certain ordinary differential equations, is computationally expensive or even infeasible. Not until recently is the notion of retractions proposed, which stimulates a significant boost in practical efficiency of Riemannian optimization methods. We refer to the monograph [AMS08] for more information on the history and the background of this subject.

On a Riemannian manifold $(\mathcal{M}, \langle\cdot,\cdot\rangle)$ equipped with the Riemannian connection ∇, let $p \in \mathcal{M}$, $R_p : T_{\mathcal{M}}(p) \to \mathcal{M}$ be a smooth mapping on a neighborhood of p, and γ be a curve on $\mathcal{M}$ such that $\gamma(t; p, v) = R_p(tv)$ is defined for any $v \in T_{\mathcal{M}}(p)$ and $t \in \mathbb{R}$ near 0. Consider the following conditions for an arbitrary $v \in T_{\mathcal{M}}(p)$:

1. $\gamma(0; p, v) = p$.

2. $\dot{\gamma}(0; p, v) = v$.

3. $\ddot{\gamma}(0; p, v) = 0$.

If the first two conditions are satisfied, then R is said to be a *(first-order) retraction* on $\mathcal{M}$ at p. If all three conditions are satisfied, then R is a *second-order retraction.* When $p \in \mathcal{M}$ also varies, the retraction R can be considered as a mapping on the tangent bundle such that $R(p, v) = R_p(v)$ for any $p \in \mathcal{M}$ and $v \in T_{\mathcal{M}}(p)$.

As we shall see in the next theorem, retractions provide a natural connection between the Riemannian gradient and Hessian and their Euclidean counterparts over the tangent space.

Theorem 4.2.2. *If $R_p : T_{\mathcal{M}}(p) \to \mathcal{M}$ is a retraction on $\mathcal{M}$ at p, then we have*

$$\mathrm{grad} f(p) = \mathrm{grad}(f \circ R_p)(0_p). \tag{4.2.4}$$

Furthermore, if the retraction R_p is of second-order, then we have

$$\mathrm{Hess} f(p) = \mathrm{Hess}(f \circ R_p)(0_p). \tag{4.2.5}$$

Proof. Our proof closely follows Proposition 5.5.4 in [AMS08].

Since the right-hand side of (4.2.4) is a Euclidean gradient, we have for an arbitrary $v \in T_{\mathcal{M}}(p)$ that

$$\begin{aligned}\langle \mathrm{grad}(f \circ R_p)(0_p), v\rangle &= \frac{d}{dt}(f \circ R_p)(tv)\Big|_{t=0} = Df(R_p(tv))\left[\frac{d}{dt}R_p(tv)\right]\Big|_{t=0} \\ &= \left\langle \mathrm{grad} f(R_p(tv)), \frac{d}{dt}R_p(tv)\right\rangle\Big|_{t=0} = \langle \mathrm{grad} f(\gamma(0; p, v)), \dot{\gamma}(0; p, v)\rangle\end{aligned}$$

$$= \langle \mathrm{grad} f(p), v \rangle.$$

The second equality in the above equation follows from the chain rule of differential maps; see Lemma 1.15 in [O'n83]. Thus, (4.2.4) is proven.

Analogously, since the right-hand side of (4.2.5) is a Euclidean Hessian, we have for an arbitrary $v \in T_{\mathcal{M}}(p)$ that

$$\begin{aligned}\langle \mathrm{Hess}(f \circ R_p)(0_p)[v], v \rangle &= \frac{d}{dt}\left(\frac{d}{dt} f(R_p(tv))\right)\Big|_{t=0} = \frac{d}{dt}\left\langle \mathrm{grad} f(R_p(tv)), \dot{\gamma}(t;p,v) \right\rangle \Big|_{t=0} \\ &= \left\langle \frac{D}{dt} \mathrm{grad} f(\gamma(t;p,v)), \dot{\gamma}(t;p,v) \right\rangle \Big|_{t=0} + \langle \mathrm{grad} f(R_p(tv)), \ddot{\gamma}(t;p,v) \rangle \Big|_{t=0} \\ &= \langle \nabla_{\dot{\gamma}(0;p,v)} \mathrm{grad} f(p), \dot{\gamma}(0;p,v) \rangle = \langle \mathrm{Hess} f(p)[v], v \rangle.\end{aligned}$$

In the above derivation, the third and fourth equalities are due to the properties of the induce covariant derivative. In view of the polarization identity, (4.2.5) is also proven. □

Riemannian Newton method

Based on what I have gathered in this section, we are now ready to transfer classical optimization algorithms over Euclidean spaces to their variants on Riemannian manifolds. Here we only focus on the simple yet classical Newton's method; see, e.g., [Smi93, EAS98, ADM$^+$02] for the early literature on Riemannian Newton method.

Let $(\mathcal{M}, \langle \cdot, \cdot \rangle)$ be a Riemannian manifold equipped with the Riemannian connection ∇. Our aim is to find a stationary point p^* of $f \in \mathfrak{F}(\mathcal{M})$ over $\mathcal{M}$, i.e. $\mathrm{grad} f(p^*) = 0$. Assume that we are supplied with a retraction R everywhere on $\mathcal{M}$.

Algorithm 4.2.3 (Riemannian Newton method)**.**
Initialize $p^0 \in \mathcal{M}$. Iterate with $k = 0, 1, 2, ...$:

1. Solve the following Newton system for $v^k \in T_{\mathcal{M}}(p^k)$:

$$\mathrm{Hess} f(p^k)[v^k] = -\mathrm{grad} f(p^k).$$

2. Set $p^{k+1} := R_{p^k}(v^k)$.

3. If the stopping criterion is not satisfied, set $k := k + 1$ and go to step 1.

The following theorem asserts that the above Riemannian Newton method attains local quadratic convergence just as for the classical Newton's method over a Euclidean space. Our proof for this theorem essentially utilizes a calculus approach in section 6.3.1 of [AMS08].

Theorem 4.2.4. *Assume that $p^k \in \mathcal{M}$ is sufficiently close to some stationary point $p^* \in \mathcal{M}$ where the Riemannian Hessian* $\mathrm{Hess} f(p^*) : T_{\mathcal{M}}(p^*) \to T_{\mathcal{M}}(p^*)$ *is nonsingular. It follows that* *the sequence $\{p^k\}$ generated by Algorithm 4.2.3 converges locally quadratically to the stationary point p^*, i.e.*

$$\limsup_{k\to\infty} \frac{\|\varphi(p^{k+1}) - \varphi(p^*)\|}{\|\varphi(p^k) - \varphi(p^*)\|^2} \leq C_q,$$

for a positive constant C_q and a local chart φ around p^*.

Proof. Consider the mapping $\phi : \mathcal{M} \to \mathcal{M}$ such that $p^{k+1} = \phi(p^k)$ is generated by the Riemannian Newton algorithm. The mapping $p^k \in \mathcal{M} \mapsto v^k \in T_{\mathcal{M}}(p^k)$ is implicitly defined through the Newton equation $\text{Hess}f(p^k)[v^k] = \nabla_{v^k}\text{grad}f(p^k) = -\text{grad}f(p^k)$. Note that in case $p^k = p^*$ we have $v^k(p^*) = 0$ and $p^{k+1} = \phi(p^*) = p^*$.

In view of Theorem 4.5.3 in [AMS08], it suffices to prove $D\phi(p^*) = 0$. By perturbing ϕ at p^* along some $w \in T_{\mathcal{M}}(p^*)$, we have

$$D\phi(p^*)[w] = D_pR(p^*, 0)[w] + D_vR(p^*, 0)[Dv(p^*)[w]] = w + Dv(p^*)[w].$$

Let us reformulate the Newton equation at p^* using the Christoffel symbol, i.e.

$$D\text{grad}f(p^*)[v(p^*)] + \Gamma_{p^*}[\text{grad}f(p^*), v(p^*)] = -\text{grad}f(p^*).$$

Perturbing p^* the above equation along w yields that

$$D\text{grad}f(p^*)[Dv(p^*)[w]] = -D\text{grad}f(p^*)[w].$$

Since $\text{Hess}f(p^*)$ is nonsingular and $\Gamma_{p^*}[\text{grad}f(p^*), \cdot] = 0$, we have that $D\text{grad}f(p^*) : T_{\mathcal{M}}(p^*) \to T_{\mathcal{M}}(p^*)$ is also nonsingular, and therefore $D\phi(p^*)[w] = 0$. Since the choice of $w \in T_{\mathcal{M}}(p^*)$ can be arbitrary, we conclude that $D\phi(p^*) = 0$ as desired. □

4.3 Robust principal component pursuit

Now we turn our attention to the problem of robust principal component pursuit. Let the observed data Z be composed in the following way:

$$Z = A_{true} + B_{true} + N,$$

where $A_{true} \in \mathcal{M}(r) = \{A \in \mathbb{R}^{m\times n} : \text{rank}(A) \leq r\}$, $B_{true} \in \mathcal{N}(s) = \{B \in \mathbb{R}^{m\times n} : \|B\|_0 \leq s\}$, and N is an m-by-n matrix of white Gaussian noise. Moreover, $\|\cdot\|_0$ denotes the number of nonzero entries of a matrix. In what follows, we omit the arguments r and s whenever their values stay constant in the context. Let the inner product $\langle\cdot,\cdot\rangle$ be defined as $\langle A, B\rangle = \text{trace}(A^\top B)$ for any $A, B \in \mathbb{R}^{m\times n}$ and $\|\cdot\|$ be the Frobenius norm. Throughout this chapter, we assume that r and s are natural numbers such that $r \ll n \leq m$ and $s \ll mn$.

Our goal is to recover the matrices A_{true} and B_{true} by solving the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & f(A, B) = \frac{1}{2}\|A + B - Z\|^2 + \frac{\mu}{2}\|A\|^2, \\ \text{subject to} \quad & (A, B) \in \mathcal{M} \times \mathcal{N}. \end{aligned} \tag{4.3.1}$$

Note that a quadratic regularization on A with $0 < \mu \ll 1$ is introduced into the objective in order to enforce the existence of solution, as provided by the following theorem.

Theorem 4.3.1. *The variational problem (4.3.1) admits a global minimizer.*

Proof. Let $(A^k, B^k) \in \mathcal{M} \times \mathcal{N}$ form an infimizing sequence for (4.3.1), i.e.

$$\lim_{k\to\infty} f(A^k, B^k) = \inf_{(A,B)\in\mathcal{M}\times\mathcal{N}} f(A, B).$$

Since f is bounded from below and coercive with respect to A and $A + B$ (i.e. $f(A, B) \to \infty$ if either $\|A\| \to \infty$ or $\|A + B\| \to \infty$), the sequences $\{A^k\}$ and $\{A^k + B^k\}$ are both uniformly bounded and, therefore, $\{B^k\}$ is also uniformly bounded. By compactness, $\{(A^k, B^k)\}$ admits an accumulation point (A^*, B^*). Moreover, note that the feasible set $\mathcal{M} \times \mathcal{N}$ is closed and $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ is continuous. Thus, we conclude that (A^*, B^*) is a global minimizer. □

Any global minimizer $(A^*, B^*) \in \mathcal{M} \times \mathcal{N}$ of (4.3.1) satisfies the first-order necessary optimality condition:

$$\begin{cases} \langle \Delta, (1+\mu)A^* + B^* - Z\rangle \geq 0, & \text{for any } \Delta \in T_{\mathcal{M}}(A^*), \\ \langle \Delta, A^* + B^* - Z\rangle \geq 0, & \text{for any } \Delta \in T_{\mathcal{N}}(B^*). \end{cases} \tag{4.3.2}$$

Here, $T_{\mathcal{M}}(A^*)$ denotes the tangent cone of the set $\mathcal{M}$ at A^*, and analogously for $T_{\mathcal{N}}(B^*)$. Note that the structure of the optimality condition (4.3.2) is due to the separability of the constraints.

Whenever $\mathrm{rank}(A^*) = r$, the set $\mathcal{M}$ is locally (around A^*) a Riemannian manifold with the Riemannian metric $\langle \cdot, \cdot \rangle$. Hence, $T_{\mathcal{M}}(A^*)$ reduces to a linear subspace in $\mathbb{R}^{m\times n}$, namely the tangent space of $\mathcal{M}$ at A^*, and the first variational inequality in (4.3.2) becomes

$$P_{T_{\mathcal{M}}(A^*)}((1+\mu)A^* + B^* - Z) = 0. \tag{4.3.3}$$

Here $P_{T_{\mathcal{M}}(A^*)}$ denotes the orthogonal projection onto the linear subspace $T_{\mathcal{M}}(A^*)$. Let $U\Sigma V^\top$ be the compact singular value decomposition (SVD) of the matrix A^*. Then the tangent space $T_{\mathcal{M}}(A^*)$ is given by

$$T_{\mathcal{M}}(A^*) = \{UMV^\top + U_pV^\top + UV_p^\top : M \in \mathbb{R}^{r\times r}, U_p \in \mathbb{R}^{m\times r}, U_p^\top U = 0, V_p \in \mathbb{R}^{n\times r}, V_p^\top V = 0\};$$

see, e.g., [Van13]. Analogously, whenever $\|B^*\|_0 = s$, $\mathcal{N}$ is a Riemannian manifold around B^* with the Riemannian metric $\langle \cdot, \cdot \rangle$. Hence, $T_{\mathcal{N}}(B^*)$ reduces to the tangent space of $\mathcal{N}$ at B^*, and correspondingly the second variational inequality in (4.3.2) becomes

$$P_{T_{\mathcal{N}}(B^*)}(A^* + B^* - Z) = 0, \tag{4.3.4}$$

where the tangent space $T_{\mathcal{N}}(B^*)$ is given by

$$T_{\mathcal{N}}(B^*) = \{\Delta \in \mathbb{R}^{m\times n} : \mathrm{supp}(\Delta) \subset \mathrm{supp}(B^*)\}.$$

4.4 Alternating minimization on matrix manifolds

In this section, we investigate the numerical solution of the variational problem (4.3.1). While (4.3.1) is handled by a rather straightforward alternating minimization scheme, the respective subproblems are sophisticated due to the respective constraint sets.

4.4.1 Alternating minimization scheme and its convergence property

We first formulate our alternating minimization scheme in Algorithm 4.4.1 below. Then a rather macroscopic convergence result for this algorithm is given in Theorem 4.4.2. While the proof for Theorem 4.4.2 is straightforward, the major work is to figure out appropriate algorithmic steps for solving the respective subproblems that activate the convergence criteria, which are the subjects of sections 4.4.2 and 4.4.3.

Algorithm 4.4.1 (Alternating minimization scheme)**.**

Initialize $A^0 \in \mathcal{M}$, $B^0 \in \mathcal{N}$. Set $k := 0$ and iterate:

1. Compute $A^{k+1} \in \mathcal{M}$ as an approximate solution for the A-subproblem: $\min_{A\in\mathcal{M}} \frac{1}{2}\|A + B^k - Z\|^2 + \frac{\mu}{2}\|A\|^2$.

2. Compute $B^{k+1} \in \mathcal{N}$ as an approximate solution for the B-subproblem: $\min_{B\in\mathcal{N}} \frac{1}{2}\|A^{k+1} + B - Z\|^2$.

3. If a suitable stopping criterion is satisfied, then stop; otherwise set $k := k + 1$ and return to step 1.

Theorem 4.4.2. *Let $\{(A^k, B^k)\} \subset \mathcal{M} \times \mathcal{N}$ be the sequence generated by Algorithm 4.4.1. Suppose that there exists a positive constant δ and two sequences of nonnegative scalars $\{\varepsilon_a^k\}$ and $\{\varepsilon_b^k\}$ such that the following conditions are satisfied for all k:*

$$f(A^{k+1}, B^k) \leq f(A^k, B^k) - \delta\|A^{k+1} - A^k\|^2, \tag{4.4.1}$$

$$f(A^{k+1}, B^{k+1}) \leq f(A^{k+1}, B^k) - \delta\|B^{k+1} - B^k\|^2, \tag{4.4.2}$$

$$\langle \Delta, (1+\mu)A^{k+1} + B^k - Z\rangle \geq -\varepsilon_a^k\|\Delta\|, \quad \text{for any } \Delta \in T_{\mathcal{M}}(A^{k+1}), \tag{4.4.3}$$

$$\langle \Delta, A^{k+1} + B^{k+1} - Z\rangle \geq -\varepsilon_b^k\|\Delta\|, \quad \text{for any } \Delta \in T_{\mathcal{N}}(B^{k+1}). \tag{4.4.4}$$

Furthermore, let $\{(A^{k^l}, B^{k^l})\}$ be any convergent subsequence of $\{(A^k, B^k)\}$ with the limit point $(A^*, B^*) \in \mathcal{M} \times \mathcal{N}$ such that $\operatorname{rank}(A^*) = r$*, $\|B^*\|_0 = s$, and* $\lim_{l\to\infty} \varepsilon_a^{k^l} = \lim_{l\to\infty} \varepsilon_b^{k^l} = 0$*. Then (A^*, B^*) satisfies the first-order optimality conditions (4.3.3)–(4.3.4).*

Proof. First note that $f(A^{k+1}, B^{k+1}) \leq f(A^{k+1}, B^k) \leq f(A^k, B^k)$ for all k. Since f is bounded from below, we have $\lim_{k\to\infty} f(A^{k+1}, B^k) - f(A^k, B^k) = \lim_{k\to\infty} f(A^{k+1}, B^k) - f(A^{k+1}, B^{k+1}) = 0$, which by conditions (4.4.1)–(4.4.2) implies that $\lim_{k\to\infty} \|A^k - A^{k+1}\| = \lim_{k\to\infty} \|B^k - B^{k+1}\| = 0$.

Now let $\{(A^{k^l}, B^{k^l})\}$ be a subsequence that converges to some (A^*, B^*) with $\operatorname{rank}(A^*) = r$ and $\|B^*\|_0 = s$. Then we have $\operatorname{rank}(A^{k^l}) = r$ and $\|B^{k^l}\|_0 = s$ for all sufficiently large l. Since $\mathcal{M}$ is a smooth manifold in a neighborhood of A^* and $\mathcal{N}$ is a smooth manifold in a neighborhood B^*, conditions (4.4.3)–(4.4.4) yield that $\|P_{T_{\mathcal{M}}(A^{k^l+1})}((1+\mu)A^{k^l+1}+B^{k^l}-Z)\| \le \varepsilon_a^{k^l}$ and $\|P_{T_{\mathcal{N}}(B^{k^l+1})}(A^{k^l+1} + B^{k^l+1} - Z)\| \le \varepsilon_b^{k^l}$ for all sufficiently large l. Due to the continuity of the mappings $(A, M) \in \mathcal{M} \times \mathbb{R}^{m\times n} \mapsto P_{T_{\mathcal{M}}(A)}(M)$ and $(B, M) \in \mathcal{N} \times \mathbb{R}^{m\times n} \mapsto P_{T_{\mathcal{N}}(B)}(M)$, we conclude that the optimality conditions (4.3.3)–(4.3.4) hold true by passing $l \to \infty$. □

In the following, we discuss in detail the resolution of the subproblems in Algorithm 4.4.1 such that the conditions (4.4.1)–(4.4.4) in Theorem 4.4.2 are fulfilled. We start by studying step 2 of Algorithm 4.4.1.

4.4.2 Sparse matrix (B-)subproblem

The global minimizer of the sparse matrix subproblem $\min_{B\in\mathcal{N}} \frac{1}{2}\|A^{k+1}+B-Z\|^2$ can be obtained explicitly in closed form by utilizing the projection operator $P_{\mathcal{N}}$. For this purpose, for a given matrix $M \in \mathbb{R}^{m\times n}$, one aligns its entries in decreasing order with respect to the absolute value; i.e. $|M_{i^1j^1}| \ge |M_{i^2j^2}| \ge ... \ge |M_{i^{mn}j^{mn}}|$. Then one obtains $P_{\mathcal{N}}(M)$ by setting

$$(P_{\mathcal{N}}(M))_{i^lj^l} = \begin{cases} M_{i^lj^l}, & \text{if } l \le s, \\ 0, & \text{otherwise.} \end{cases}$$

Note that $P_{\mathcal{N}}(M)$ is not unique if $M_{i^sj^s} = M_{i^{s+1}j^{s+1}}$. In this case we simply take $P_{\mathcal{N}}(M)$ to be any one of the valid candidates. With $P_{\mathcal{N}}$ at hand, the global minimizer of the sparse matrix subproblem is computed as in step 1 of Algorithm 4.4.3 below. On the other hand, such a global minimizer does not necessarily guarantee a sufficient decrease in the objective as required by condition (4.4.2). When the global minimizer fails to fulfill condition (4.4.2), we resort to a local minimizer as specified by step 3 in Algorithm 4.4.3.

Algorithm 4.4.3 (B-subproblem solver).
Let $(A^{k+1}, B^k) \in \mathcal{M} \times \mathcal{N}$ be given. Choose $0 < \delta \le 1/2$.

1. Compute the global minimizer of the B-subproblem $\widehat{B}^{k+1} = P_{\mathcal{N}}(Z - A^{k+1})$.
2. If $f(A^{k+1}, \widehat{B}^{k+1}) \le f(A^{k+1}, B^k) - \delta\|\widehat{B}^{k+1} - B^k\|^2$, then accept $B^{k+1} = \widehat{B}^{k+1}$; otherwise reject $\widehat{B}^{k+1}$ and continue with step 3.
3. Return B^{k+1} with

$$B_{ij}^{k+1} = \begin{cases} (Z - A^{k+1})_{ij}, & \text{if } B_{ij}^k \ne 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.4.5}$$

Theorem 4.4.4. *The solution B^{k+1} computed by Algorithm 4.4.3 satisfies condition (4.4.2). Moreover, if $\|B^{k+1}\|_0 = s$, condition (4.4.4) holds with $\varepsilon_b^k = 0$.*

Proof. (Case 1). We first prove the conclusion in the case that $\widehat{B}^{k+1}$ is accepted. It follows immediately from step 2 of Algorithm 4.4.3 that condition (4.4.2) holds.

Now assume that $\|B^{k+1}\|_0 = s$, which implies that the tangent space $T_{\mathcal{N}}(B^{k+1}) = \{\Delta \in \mathbb{R}^{m\times n} : \text{supp}(\Delta) \subset \text{supp}(B^{k+1})\}$. Then it follows that $\Delta_{ij} = 0$ whenever $(i,j) \notin \text{supp}(B^{k+1})$ and that $B_{ij}^{k+1} = (Z - A^{k+1})_{ij}$ whenever $(i,j) \in \text{supp}(B^{k+1})$. Therefore, $\langle \Delta, A^{k+1} + B^{k+1} - Z\rangle = 0$ for any $\Delta \in T_{\mathcal{N}}(B^{k+1})$ and (4.4.4) holds with $\varepsilon_b^k = 0$.

(Case 2). Now consider the case where $\widehat{B}^{k+1}$ is not accepted. Then (4.4.5) must hold true; i.e. we have that $B_{ij}^{k+1} = (Z - A^{k+1})_{ij}$ whenever $(i,j) \in \text{supp}(B^k)$ and that $(B^k - B^{k+1})_{ij} = 0$ whenever $(i,j) \notin \text{supp}(B^k)$. Thus condition (4.4.2) is fulfilled since

$$\begin{aligned} f(A^{k+1}, B^k) - f(A^{k+1}, B^{k+1}) &= \frac{1}{2}\|B^k - B^{k+1}\|^2 + \langle B^k - B^{k+1}, A^{k+1} + B^{k+1} - Z\rangle \\ &= \frac{1}{2}\|B^k - B^{k+1}\|^2 \geq \delta \|B^k - B^{k+1}\|^2. \end{aligned}$$

The argument for the satisfaction of condition (4.4.4) with $\varepsilon_b^k = 0$ is analogous to the one given in Case 1. □

In the numerical implementation of step 1 of Algorithm 4.4.3, we call the Matlab command `sort`, which is based on a Quicksort algorithm of complexity $O(mn\log(mn))$ in average. We remark that, according to our numerical experience, the overall cost of the alternating minimization scheme is dominated by the A-subproblem solve. Therefore, in this work we do not pursue more advanced randomized partial ordering algorithms [Knu97], e.g. Quickselect, for further CPU gain. We also remark that the choice of the parameter δ in Algorithm 4.4.3 represents a tradeoff between the convergence of the iterates and the global optimality of their limit. In fact, if δ is too large, then the iterates may possibly converge to one among many undesired local solutions. On the other hand, it may slow down the speed of convergence by choosing δ too close to 0.

4.4.3 Low-rank matrix (A-)subproblem

Now we turn our attention to solving the A-subproblem, namely the task of step 1 of Algorithm 4.4.1. Unlike solving the B-subproblem in section 4.4.2, the A-subproblem will be resolved inexactly by a single update of a gradient-based algorithm. The organization of this subsection is as follows. In this subsection, we first review the global minimizer via SVD and discuss its drawbacks in numerical computation. We then develop a projected dogleg method on the fixed-rank matrix manifold for resolving the A-subproblem and its convergence property is studied in detail.

Global minimizer via (partial) SVD

The low-rank matrix subproblem $\min_{A\in\mathcal{M}} \frac{1}{2}\|A + B^k - Z\|^2 + \frac{\mu}{2}\|A\|^2$ admits a global minimizer in closed form for any $\mu \geq 0$. It is obtained by the projection of $\frac{1}{1+\mu}(Z - B^k)$ onto $\mathcal{M}$, denoted by $P_{\mathcal{M}}(\frac{1}{1+\mu}(Z - B^k))$. Let $U_z^k \Sigma_z^k (V_z^k)^\top$ be the singular-value decomposition (SVD) of the matrix $\frac{1}{1+\mu}(Z - B^k)$, where $U_z^k \in \mathbb{R}^{m\times m}$ and $V_z^k \in \mathbb{R}^{n\times n}$ are both orthogonal matrices, and Σ_z^k is a diagonal matrix in $\mathbb{R}^{m\times n}$ with nonnegative diagonal elements $(\sigma_z^k)_1,\ (\sigma_z^k)_2,\ ...,\ (\sigma_z^k)_m$ in decreasing order. Then, by the well-known Eckart-Young theorem [EY36], a global minimizer of the A-subproblem is given by $P_{\mathcal{M}}(Z - B^k) = U_z^k \widehat{\Sigma}_z^k (V_z^k)^\top$, where $\widehat{\Sigma}_z^k$ is a diagonal matrix in $\mathbb{R}^{m\times n}$ with diagonal elements $(\sigma_z^k)_1,\ (\sigma_z^k)_2,\ ...,\ (\sigma_z^k)_r,\ 0, ...,\ 0$.

The classical SVD of an m-by-n matrix has a complexity of $O(mn^2)$ flops [TB97], which is rather expensive in large-scale computation. In the context of our low-rank subproblem, however, this can be accelerated by a partial SVD technique of complexity $O(mnr)$, see e.g. the package PROPACK [PRO] available online, since only the first r-th singular values and vectors are needed. Although such a global minimization strategy often works quite efficiently, see the corresponding numerical tests in section 4.5, it does not guarantee for the overall alternating minimization scheme (global) convergence towards a stationary point from an arbitrary initial guess. In particular, satisfaction of the sufficient conditions for global convergence, i.e. conditions (4.4.1) and (4.4.3), is not ensured. For this sake, in the following we investigate in detail an inexact-solution strategy for the low-rank subproblem based on Riemannian optimization, which fulfills conditions (4.4.1) and (4.4.3), thus admitting a global convergence theory for Algorithm 4.4.1. The proposed method also enjoys good practical efficiency as will be demonstrated in section 4.5.

Projected dogleg method on a fixed-rank matrix manifold

Riemannian optimization techniques have been developed in the past two decades; see, e.g., [Smi93, EAS98]. More recently, these methods have been successfully applied to optimization problems related to low-rank matrices [SE10, BA11, Van13]. In the following, we develop a tailored Riemannian optimization approach, namely a projected dogleg method, on a rank-r matrix manifold $\bar{\mathcal{M}}(r) = \{A \in \mathbb{R}^{m\times n} : \operatorname{rank}(A) = r\}$. We emphasize that the ultimate goal of the projected dogleg method under consideration is to fulfill conditions (4.4.1) and (4.4.3) with $\lim_{k\to\infty} \varepsilon_a^k = 0$ in order to guarantee the global convergence of the alternating minimization scheme. Other Riemannian approaches, e.g. Riemannian trust-region method [AMS08, BMAS14], may also be applicable in the context, but require a rather different, perhaps more involved, analysis.

Given $B^k \in \mathcal{N}$, define the smooth mapping $f_A^k : \bar{\mathcal{M}} \to \mathbb{R}$ with $f_A^k(A) = f(A, B^k)$ for all $A \in \bar{\mathcal{M}}$. The Riemannian gradient of f_A^k at A on $\bar{\mathcal{M}}$, denoted by $\operatorname{grad} f_A^k(A)$, is defined as a tangent vector in the tangent space $T_{\bar{\mathcal{M}}}(A)$ such that $\langle \operatorname{grad} f_A^k(A), \Delta\rangle = Df_A^k(A)[\Delta]$ for all $\Delta \in T_{\bar{\mathcal{M}}}(A)$. Here $Df_A^k(A)[\Delta]$ is the directional derivative of f_A^k at A along the direction Δ. Let ∇ be the (unique) Riemannian connection on $\bar{\mathcal{M}}$, and let $\nabla_{\eta(A)}\xi(A) \in T_{\bar{\mathcal{M}}}(A)$ denote the covariant derivative of two smooth vector fields ξ and η on $\bar{\mathcal{M}}$ at A. Then the Riemannian Hessian of f_A^k at A on $\bar{\mathcal{M}}$, denoted by $\operatorname{Hess} f_A^k(A)$, is a linear mapping from $T_{\bar{\mathcal{M}}}(A)$ to $T_{\bar{\mathcal{M}}}(A)$ such that $\operatorname{Hess} f_A^k(A)[\Delta] = \nabla_\Delta \operatorname{grad} f_A^k(A)$ for any $\Delta \in T_{\bar{\mathcal{M}}}(A)$. By considering $\bar{\mathcal{M}}$ as an embedded submanifold in the Euclidean space $(\mathbb{R}^{m\times n}, \langle\cdot,\cdot\rangle)$, the Riemannian gradient is derived as

$$\mathrm{grad} f_A^k(A) = P_{T_{\bar{\mathcal{M}}}(A)}(\nabla f_A^k(A)) = P_{T_{\bar{\mathcal{M}}}(A)}((1+\mu)A + B^k - Z),$$

see section 3.6.1 in [AMS08]. The derivation of the Riemannian Hessian is more involved in general. For the rank-r matrix manifold, the following Hessian formula can be calculated by constructing a factorization-based second-order retraction [Van13]:

$$\begin{aligned}\mathrm{Hess} f_A^k(A)[\Delta] &= (1+\mu)\Delta + (I - UU^\top)\nabla f_A^k(A)(I - VV^\top)\Delta^\top U\Sigma^{-1}V^\top \\ &\quad + U\Sigma^{-1}V^\top\Delta^\top(I - UU^\top)\nabla f_A^k(A)(I - VV^\top) \\ &= (1+\mu)\Delta + (I - UU^\top)(B^k - Z)(I - VV^\top)\Delta^\top U\Sigma^{-1}V^\top \\ &\quad + U\Sigma^{-1}V^\top\Delta^\top(I - UU^\top)(B^k - Z)(I - VV^\top),\end{aligned} \tag{4.4.6}$$

where $A = U\Sigma V^\top$ is the compact SVD of A with a full-rank diagonal matrix $\Sigma \in \mathbb{R}^{r\times r}$ and two orthonormal matrices $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$. It is worth noting that the Hessian formula (4.4.6) should be handled in a matrix-free fashion so that computing each matrix-vector product $\mathrm{Hess} f_A^k(A)[\Delta]$ requires $O(mnr)$ flops. To ease our presentation, in the remainder of section 4.4.3, we use the notations $g^k := \mathrm{grad} f_A^k(A^k)$, $H^k := \mathrm{Hess} f_A^k(A^k)$, and assume that $g^k \neq 0$.

One can approximate f_A^k around A^k in the tangent space $T_{\bar{\mathcal{M}}}(A^k)$ by a quadratic function $h^k(\Delta^k) := f_A^k(A^k) + \langle g^k, \Delta^k\rangle + \frac{1}{2}\langle \Delta^k, H^k[\Delta^k]\rangle$ for $\Delta^k \in T_{\bar{\mathcal{M}}}(A^k)$. Presuming that H^k is positive definite on $T_{\bar{\mathcal{M}}}(A^k)$, based on the Cauchy point

$$\Delta_C^k := -\frac{\|g^k\|^2}{\langle g^k, H^k[g^k]\rangle} g^k, \tag{4.4.7}$$

and the Newton point

$$\Delta_N^k := -(H^k)^{-1}[g^k], \tag{4.4.8}$$

we define the dogleg path in the tangent space $T_{\bar{\mathcal{M}}}(A^k)$ as follows:

$$\Delta^k(\tau^k) = \begin{cases} \tau^k \Delta_C^k, & \text{if } 0 \le \tau^k \le 1, \\ \Delta_C^k + (\tau^k - 1)(\Delta_N^k - \Delta_C^k), & \text{if } 1 \le \tau^k \le 2. \end{cases} \tag{4.4.9}$$

Lemma 4.4.5. *For the statements:*

i. H^k is positive definite on $T_{\bar{\mathcal{M}}}(A^k)$; i.e. $\langle \Delta, H^k[\Delta]\rangle > 0$ for any nonzero $\Delta \in T_{\bar{\mathcal{M}}}(A^k)$;

ii. $\langle \Delta_C^k, \Delta_N^k - \Delta_C^k\rangle \ge 0$;

iii. $\|\Delta^k(\tau^k)\|$ is an increasing function in $\tau^k \in [0,2]$;

the following implication holds true: (i) ⇒ (ii) ⇒ (iii).

Proof. The proof is analogous to the one of Lemma 4.2 in [NW06]. □

Given the current iterate $A^k \in \bar{\mathcal{M}}$, in order to generate the next iterate from the update step in the tangent space at A^k, we use the metric projection $P_{\bar{\mathcal{M}}} : \mathbb{R}^{m\times n} \to \bar{\mathcal{M}}$ defined by $P_{\bar{\mathcal{M}}}(Z) = \arg\min_{A\in\bar{\mathcal{M}}} \|A - Z\|$, which makes a smooth mapping locally around A^k; see, e.g., [LM08]. Different from the scenario in section 4.4.3, given $A \in \bar{\mathcal{M}}$ and $\Delta \in T_{\bar{\mathcal{M}}}(A)$, the projection $P_{\bar{\mathcal{M}}}(A + \Delta)$ can be computed via a reduced SVD on a $2r$-by-$2r$ matrix thanks to unitary invariance; see, e.g., [NS12, Van13]. The reduction of the computational cost, compared to the approach in section 4.4.3, is significant in practice where r is typically much smaller than m and n. For the reader's convenience, we describe the implementation of the projection operation in Algorithm 4.4.6 below.

Algorithm 4.4.6 (Projection onto fixed-rank matrix manifold via reduced SVD)**.**

Let $A \in \bar{\mathcal{M}}(r)$, represented in the compact SVD form $A = U\Sigma V^\top$, and $\Delta \in T_{\bar{\mathcal{M}}(r)}(A)$ be given. Choose $0 < \epsilon_s \ll 1$.

1. Compute $M = U^\top \Delta V$, $U_p = \Delta V - UM$, $V_p = \Delta^\top U - VM^\top$.
2. Perform the QR-factorization of U_p and V_p such that $U_p = Q_u R_u$ and $V_p = Q_v R_v$ with two orthonormal matrices $Q_u \in \mathbb{R}^{m\times r}$, $Q_v \in \mathbb{R}^{n\times r}$ and two upper-triangular matrices $R_u, R_v \in \mathbb{R}^{r\times r}$.
3. Perform an SVD of the $2r$-by-$2r$ matrix on the left-hand side of the following equation:
$$\begin{bmatrix} \Sigma + M & R_v^\top \\ R_u & 0 \end{bmatrix} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top,$$
where $\widetilde{\Sigma} = \mathrm{diag}(\{\widetilde{\sigma}_j\}_{j=1}^{2r}) \in \mathbb{R}^{2r\times 2r}$ is some diagonal matrix with positive diagonal entries $\{\widetilde{\sigma}_j\}_{j=1}^{2r}$ in descending order, and $\widetilde{U}, \widetilde{V} \in \mathbb{R}^{2r\times 2r}$ are two orthogonal matrices.
4. Set $\widehat{\Sigma} = \mathrm{diag}(\{\max(\widetilde{\sigma}_j, \epsilon_s)\}_{j=1}^{r}) \in \mathbb{R}^{r\times r}$, $\widehat{U} = [U\ Q_u][\{\widetilde{U}_j\}_{j=1}^{r}] \in \mathbb{R}^{m\times r}$, and $\widehat{V} = [V\ Q_v][\{\widetilde{V}_j\}_{j=1}^{r}] \in \mathbb{R}^{n\times r}$, where $\widetilde{U}_j$ and $\widetilde{V}_j$ denote the j-th columns of $\widetilde{U}$ and $\widetilde{V}$, respectively. Return $P_{\bar{\mathcal{M}}(r)}(A + \Delta) = \widehat{U}\widehat{\Sigma}\widehat{V}^\top$.

Concerning the QR-factorization of the matrices U_p and V_p required in step 2 of Algorithm 4.4.6, we remark that in a MATLAB environment one may call the command `qr` with the "economy-size" option. In addition, we note that a small positive parameter ϵ_s is introduced in step 4 in order to prevent rank deficiency of the projection. Ideally, it suffices to choose $\epsilon_s > 0$ which is significantly smaller than the minimal nonzero singular value of the underlying low-rank matrix A^* that, together with B^*, solves (4.3.1). Throughout our numerical experiments in section 4.5, we shall fix $\epsilon_s = 10^{-3}$. For a proper tuning of the underlying rank r along the overall iterative algorithm, we refer to the trimming procedure presented in section 4.4.4.

It is known [LM08, AM12] that for any point A on the smooth manifold $\bar{\mathcal{M}}$, the projection $P_{\bar{\mathcal{M}}}$ is a smooth diffeomorphism in a neighborhood of A, and moreover the differentiation rule

$$DP_{\bar{\mathcal{M}}}(A)[\Delta] = P_{T_{\bar{\mathcal{M}}}(A)}(\Delta)$$

holds for any $\Delta \in \mathbb{R}^{m\times n}$. Thus, the projected dogleg path $\tau^k \mapsto P_{\bar{\mathcal{M}}}(A^k + \Delta^k(\tau^k))$ is a well-defined smooth function in a neighborhood of 0. We remark that, in the context of [AMS08, AM12], $P_{\bar{\mathcal{M}}}$ induces a second-order retraction on $\bar{\mathcal{M}}$ near A given by $\Delta \in T_{\bar{\mathcal{M}}}(A) \mapsto P_{\bar{\mathcal{M}}}(A + \Delta) \in \bar{\mathcal{M}}$, which locally fits the exponential mapping up to second order.

We are now in a position to present the projected dogleg method for solving the low-rank matrix subproblem. Below, we have chosen a specific sequence of trial step sizes $\mathcal{F}^k$, but obviously other choices may be considered as well.

Algorithm 4.4.7 (A-subproblem solver via projected dogleg method)**.**

Let $(A^k, B^k) \in \bar{\mathcal{M}} \times \mathcal{N}$ be given. Choose $\delta > 0$.

0.[#] (Optional) Compute the global minimizer of the A-subproblem $\widehat{A}^{k+1} = P_{\mathcal{M}}(\frac{1}{1+\mu}(Z - B^k))$. If $f_A^k(\widehat{A}^{k+1}) \le f_A^k(A^k) - \delta\|\widehat{A}^{k+1} - A^k\|^2$, then accept $A^{k+1} = \widehat{A}^{k+1}$; otherwise, reject $\widehat{A}^{k+1}$ and continue with step 1.

1. Compute g^k, H^k. If $\langle g^k, H^k[g^k]\rangle > 0$, then compute Δ_C^k by formula (4.4.7); otherwise, set $\Delta^k(\tau^k) := -\tau^k g^k$, $\mathcal{F}^k := \{1, 1/2, 1/4, 1/8, 1/16, ...\}$, and go to step 3.

2. Compute Δ_N^k by formula (4.4.8). If $\langle \Delta_C^k, \Delta_N^k - \Delta_C^k\rangle < 0$ or any non-positive definiteness of H^k is detected during the computation, then set $\Delta^k(\tau^k) := \tau^k\Delta_C^k$ and $\mathcal{F}^k := \{1, 1/2, 1/4, 1/8, 1/16, ...\}$; otherwise define the dogleg path $\Delta^k : [0, 2] \to T_{\bar{\mathcal{M}}}(A^k)$ as in (4.4.9) and set $\mathcal{F}^k := \{2, 3/2, 1, 1/2, 1/4, 1/8, 1/16, ...\}$.

3. Set τ^k to be the largest element in $\mathcal{F}^k$ that fulfills
$$f_A^k(A^k) - f_A^k(P_{\bar{\mathcal{M}}}(A^k + \Delta^k(\tau^k))) \ge \delta\|A^k - P_{\bar{\mathcal{M}}}(A^k + \Delta^k(\tau^k))\|^2. \tag{4.4.10}$$
Return $A^{k+1} = P_{\bar{\mathcal{M}}}(A^k + \Delta^k(\tau^k))$.

We remark that step 0 is included in the above algorithm as an optional trial step, only recommended for utility when the global minimizer $\widehat{A}^{k+1}$ tends to be accepted and can be computed at low cost, say e.g. via partial SVD [PRO]. Since the projected dogleg method works practically well in its own right, unless otherwise specified, this trial step is skipped in our subsequent algorithmic development and analysis. Nevertheless, in section 4.5.2 we shall numerically compare the performances of Algorithm 4.4.7 both with and without step 0, together with the augmented Lagrangian method based on a convex variational model.

Lemma 4.4.8. *There exists $\bar{\tau}^k > 0$ such that condition (4.4.10) is fulfilled for all $\tau^k \in (0, \bar{\tau}^k]$. Consequently, step 3 in Algorithm 4.4.7 always returns some admissible step size $\tau^k > 0$ fulfilling condition (4.4.1) after finitely many trials.*

[#] This trial step is optional, which is only recommended for utility if the global minimizer tends to be accepted and can be computed at low cost. Unless otherwise specified, this step is skipped in our algorithmic development and analysis.

Proof. Let $\phi(\tau^k) := f_A^k(A^k) - f_A^k(P_{\bar{\mathcal{M}}}(A^k + \Delta^k(\tau^k))) - \delta\|A^k - P_{\bar{\mathcal{M}}}(A^k + \Delta^k(\tau^k))\|^2$, which is a well-defined smooth function in a neighborhood of 0. Then it follows that $\phi(0) = 0$ and

$$\phi'(0) = -\langle g^k, (\Delta^k)'(0)\rangle \geq \min\left(1, \frac{\|g^k\|^2}{\langle g^k, H^k[g^k]\rangle}\right)\|g^k\|^2 \geq \min\left(1, \frac{1}{\lambda_{\max}(H^k)}\right)\|g^k\|^2 > 0.$$

Since ϕ is continuously differentiable in a neighborhood of 0, there exists some $\bar{\tau}^k > 0$ such that $\phi'(\cdot) > 0$ on the interval $(0, \bar{\tau}^k]$. By utilizing the mean value theorem, we conclude that $\phi(\cdot) \geq 0$ on the interval $(0, \bar{\tau}^k]$. □

Lemma 4.4.9. *Let $\{A^k\} \subset \bar{\mathcal{M}}$ be generated by Algorithm 4.4.7 along with some sequence $\{B^k\} \subset \mathcal{N}$ satisfying condition (4.4.2). Then the following statements hold true:*

i. $\lim_{k\to\infty} \|A^k - A^{k+1}\| = 0$.

ii. $\lim_{k\to\infty} \|\Delta^k(\tau^k))\| = 0$.

iii. Any convergent subsequence $\{A^{k^l}\}$ of $\{A^k\}$ satisfies $\lim_{l\to\infty} \|g^{k^l}\| = 0$.

Proof. Owing to Lemma 4.4.8, the proof of (i) essentially resembles the first part of the proof for Theorem 4.4.2.

Concerning (ii), note that $A^{k+1} = P_{\bar{\mathcal{M}}}(A^k + \Delta^k(\tau^k))$, which satisfies the necessary condition $P_{T_{\bar{\mathcal{M}}}(A^{k+1})}(A^{k+1} - A^k - \Delta^k(\tau^k)) = 0$. Then it follows from the reverse triangle inequality that

$$\|P_{T_{\bar{\mathcal{M}}}(A^{k+1})}(\Delta^k(\tau^k))\| \leq \|P_{T_{\bar{\mathcal{M}}}(A^{k+1})}(A^{k+1} - A^k - \Delta^k(\tau^k))\| + \|A^k - A^{k+1}\| \to 0,$$

and therefore

$$\begin{aligned}\|\Delta^k(\tau^k)\| &= \|P_{T_{\bar{\mathcal{M}}}(A^k)}(\Delta^k(\tau^k))\| \\ &\leq \|P_{T_{\bar{\mathcal{M}}}(A^{k+1})}(\Delta^k(\tau^k))\| + \|P_{T_{\bar{\mathcal{M}}}(A^k)}(\Delta^k(\tau^k)) - P_{T_{\bar{\mathcal{M}}}(A^{k+1})}(\Delta^k(\tau^k))\| \to 0,\end{aligned}$$

as $k \to \infty$.

We prove (iii) by contradiction. For this purpose, let $\{A^{k^l}\}$ be a convergent subsequence of $\{A^k\}$ and $\varepsilon > 0$ such that $\|g^{k^l}\| \geq \varepsilon$ for all l. Based on an observation of the structure of the Riemannian Hessian given in (4.4.6), the sequence $\{H^{k^l}\}$ is uniformly bounded, and we denote $\kappa_h := \sup_l \lambda_{\max}(H^{k^l})$. Making use of Lemma 4.4.5(iii), we obtain a lower bound for $\|\Delta^{k^l}(\tau^{k^l})\|$ as follows:

$$\|\Delta^{k^l}(\tau^{k^l})\| \geq \min\left(\tau^{k^l}\|g^{k^l}\|, \min(1, \tau^{k^l})\frac{\|g^{k^l}\|^3}{\langle g^{k^l}, H^{k^l}[g^{k^l}]\rangle}\right) \geq \varepsilon \min(1, 1/\kappa_h)\min(1, \tau^{k^l}).$$

Then the result in (ii) yields that $\lim_{l\to\infty} \tau^{k^l} = 0$. Due to the nature of the backtracking dogleg search in step 3 of Algorithm 4.4.7, this further implies that the trial step $2\tau^{k^l}$ is not admissible at iteration l for all sufficiently large l; i.e.

$$f_A^{k^l}(A^{k^l}) - f_A^{k^l}(P_{\bar{\mathcal{M}}}(A^{k^l} + \Delta^{k^l}(2\tau^{k^l}))) < \delta\|A^{k^l} - P_{\bar{\mathcal{M}}}(A^{k^l} + \Delta^{k^l}(2\tau^{k^l}))\|^2,$$

for all l. Dividing both sides above by $2\tau^{k^l}$ and passing $l \to \infty$, we find

$$0 \geq \lim_{l\to\infty} -\langle g^{k^l}, (\Delta^{k^l})'(0)\rangle \geq \min\left(1, \frac{1}{\kappa_h}\right) \lim_{l\to\infty} \|g^{k^l}\|^2,$$

which leads to a contradiction. □

Lemma 4.4.10. *Let $(A^*, B^*) \in \bar{\mathcal{M}} \times \mathcal{N}$ satisfy the first-order optimality conditions (4.3.3)–(4.3.4) with $\|B^*\|_0 = s$. Further assume that the Riemannian Hessian $\mathrm{Hess} f(A^*, B^*) : T_{\bar{\mathcal{M}}}(A^*) \times T_{\mathcal{N}}(B^*) \to T_{\bar{\mathcal{M}}}(A^*) \times T_{\mathcal{N}}(B^*)$ is strictly positive definite when $\mu = 0$. Then the following statements hold true:*

i. The Riemannian Hessian of f with respect to the first argument, denoted by $\mathrm{Hess}_A f(A^*, B^*) : T_{\bar{\mathcal{M}}}(A^*) \to T_{\bar{\mathcal{M}}}(A^*)$, is strictly positive definite for any $\mu \geq 0$.

ii. The following tangent space transversality holds true:

$$T_{\bar{\mathcal{M}}}(A^*) \cap T_{\mathcal{N}}(B^*) = \{0\}. \tag{4.4.11}$$

iii. The linear operator $P_{T_{\bar{\mathcal{M}}}(A^*)} \circ P_{T_{\mathcal{N}}(B^*)} : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is a contraction; i.e. there exists a constant $\kappa_p \in [0, 1)$ such that

$$\|(P_{T_{\bar{\mathcal{M}}}(A^*)} \circ P_{T_{\mathcal{N}}(B^*)})(\Delta)\| \leq \kappa_p \|\Delta\|, \tag{4.4.12}$$

for all $\Delta \in \mathbb{R}^{m\times n}$.

Proof. Given an arbitrary nonzero element Δ_A in $T_{\bar{\mathcal{M}}}(A^*)$, we have

$$\begin{aligned}
0 &< \langle (\Delta_A, 0), \mathrm{Hess} f(A^*, B^*)[(\Delta_A, 0)]\rangle_{T_{\bar{\mathcal{M}}}(A^*)\times T_{\mathcal{N}}(B^*)}\Big|_{\mu=0} \\
&= \langle (\Delta_A, 0), \nabla_{(\Delta_A,0)} \mathrm{grad} f(A^*, B^*)\rangle_{T_{\bar{\mathcal{M}}}(A^*)\times T_{\mathcal{N}}(B^*)}\Big|_{\mu=0} \\
&= \langle \Delta_A, \nabla_{\Delta_A} \mathrm{grad}_A f(A^*, B^*)\rangle_{T_{\bar{\mathcal{M}}}(A^*)}\Big|_{\mu=0} \\
&= \langle \Delta_A, \mathrm{Hess}_A f(A^*, B^*)[\Delta_A]\rangle_{T_{\bar{\mathcal{M}}}(A^*)}\Big|_{\mu=0} \\
&\leq \langle \Delta_A, \mathrm{Hess}_A f(A^*, B^*)[\Delta_A]\rangle_{T_{\bar{\mathcal{M}}}(A^*)}\Big|_{\mu\geq 0}.
\end{aligned}$$

The last inequality follows from an observation of the Hessian formula (4.4.6). Thus, (i) is proven.

We prove (ii) by contradiction. For this purpose, assume that $\mu = 0$ and there exists a nonzero element $\Delta \in T_{\bar{\mathcal{M}}}(A^*) \cap T_{\mathcal{N}}(B^*)$. Since $\bar{\mathcal{M}}$ is an embedded submanifold of $\mathbb{R}^{m\times n}$, we have

$$\begin{aligned}
&\mathrm{Hess} f(A^*, B^*)[(\Delta, -\Delta)] \\
&= P_{T_{\bar{\mathcal{M}}}(A^*)\times T_{\mathcal{N}}(B^*)} \left(D_{(A,B)} \mathrm{grad} f(A^*, B^*)[(\Delta, -\Delta)]\right)
\end{aligned}$$

$$
\begin{aligned}
&= P_{T_{\bar{\mathcal{M}}}(A^*)\times T_{\mathcal{N}}(B^*)}\left(D_{(A,B)}\left(P_{T_{\bar{\mathcal{M}}}(A)}(A+B-Z), P_{T_{\mathcal{N}}(B)}(A+B-Z)\right)[(\Delta,-\Delta)]\Big|_{(A,B)=(A^*,B^*)}\right)\\
&= (0,0). \qquad (4.4.13)
\end{aligned}
$$

In the above formulae, note that the first equality follows from Proposition 5.3.2 in [AMS08]. The last equality is a consequence of conditions (4.3.3)–(4.3.4) and the chain rule of differentiation. Thus, (4.4.13) yields $\langle(\Delta,-\Delta),\mathrm{Hess}f(A^*,B^*)[(\Delta,-\Delta)]\rangle = 0$, which contradicts the positive definiteness of $\mathrm{Hess}f(A^*,B^*)$.

We prove (iii) again by contradiction. Note that the composition of projections $P_{T_{\bar{\mathcal{M}}}(A^*)} \circ P_{T_{\mathcal{N}}(B^*)}$ is nonexpansive, and therefore condition (4.4.12) always hold true for $\kappa_p = 1$. Now assume that (iii) does not hold. Then there must exist a sequence $\{\kappa_p^l\}_{l=1}^{\infty} \in [0,1)$ with $\lim_{l\to\infty}\kappa_p^l = 1$ and correspondingly $\{\Delta^l\}\subset\mathbb{R}^{m\times n}$ with $\|\Delta^l\| = 1$ for all l such that it holds for all l that

$$
\kappa_p^l\|\Delta^l\| < \|(P_{T_{\bar{\mathcal{M}}}(A^*)}\circ P_{T_{\mathcal{N}}(B^*)})(\Delta^l)\| \le \|\Delta^l\|.
$$

Upon the extraction of a subsequence of $\{\Delta^l\}$, whose limit point $\Delta\in\mathbb{R}^{m\times n}$ satisfies $\|\Delta\| = 1$, we have

$$
\|(P_{T_{\bar{\mathcal{M}}}(A^*)}\circ P_{T_{\mathcal{N}}(B^*)})(\Delta)\| = \|\Delta\|.
$$

Then it follows from the self-adjointness and idempotence of an orthogonal projection onto a linear subspace that

$$
\begin{aligned}
\langle\Delta,\Delta\rangle &= \langle(P_{T_{\bar{\mathcal{M}}}(A^*)}\circ P_{T_{\mathcal{N}}(B^*)})(\Delta),(P_{T_{\bar{\mathcal{M}}}(A^*)}\circ P_{T_{\mathcal{N}}(B^*)})(\Delta)\rangle\\
&= \langle\Delta,(P_{T_{\mathcal{N}}(B^*)}\circ P_{T_{\bar{\mathcal{M}}}(A^*)}\circ P_{T_{\mathcal{N}}(B^*)})(\Delta)\rangle,
\end{aligned}
$$

or equivalently $\langle\Delta,(\mathrm{id}-P_{T_{\mathcal{N}}(B^*)}\circ P_{T_{\bar{\mathcal{M}}}(A^*)}\circ P_{T_{\mathcal{N}}(B^*)})(\Delta))\rangle = 0$. By the self-adjointness we have $(\mathrm{id}-P_{T_{\mathcal{N}}(B^*)}\circ P_{T_{\bar{\mathcal{M}}}(A^*)}\circ P_{T_{\mathcal{N}}(B^*)})(\Delta) = 0$, and thus

$$
(P_{T_{\mathcal{N}}(B^*)}\circ P_{T_{\bar{\mathcal{M}}}(A^*)}\circ P_{T_{\mathcal{N}}(B^*)})(\Delta) = P_{T_{\mathcal{N}}(B^*)}(\Delta). \qquad (4.4.14)
$$

In particular, note that $P_{T_{\mathcal{N}}(B^*)}(\Delta)\in T_{\mathcal{N}}(B^*)$ and $P_{T_{\mathcal{N}}(B^*)}(\Delta)\neq 0$. Further manipulation of (4.4.14) yields that $\langle P_{T_{\mathcal{N}}(B^*)}(\Delta),(\mathrm{id}-P_{T_{\bar{\mathcal{M}}}(A^*)})(P_{T_{\mathcal{N}}(B^*)}(\Delta))\rangle = 0$, and therefore by the self-adjointness

$$
P_{T_{\mathcal{N}}(B^*)}(\Delta) = (P_{T_{\bar{\mathcal{M}}}(A^*)}\circ P_{T_{\mathcal{N}}(B^*)})(\Delta).
$$

Thus, we have found $P_{T_{\mathcal{N}}(B^*)}(\Delta)\neq 0$ such that $P_{T_{\mathcal{N}}(B^*)}(\Delta)\in T_{\bar{\mathcal{M}}}(A^*)\cap T_{\mathcal{N}}(B^*)$, which contradicts the tangent space transversality in (ii). □

We remark that in [CSPW11] the tangent space transversality condition (4.4.11) is discussed in detail, and a sufficient condition on A^* and B^*, which holds with high probability in practice, is also provided for ensuring the tangent space transversality. From Lemma 4.4.10, we see that tangent space transversality can be naturally regarded as a consequence of the second-order sufficient optimality condition.

Theorem 4.4.11. *Let $\{A^k\} \subset \bar{\mathcal{M}}$ be a sequence generated by Algorithm 4.4.7 along with some sequence $\{B^k\} \subset \mathcal{N}$ generated by Algorithm 4.4.3. At iteration k, assume that the iterate (A^k, B^k) is sufficiently close to some $(A^*, B^*) \in \bar{\mathcal{M}} \times \mathcal{N}$ with $\|B^*\|_0 = s$ satisfying the first-order optimality conditions (4.3.3)–(4.3.4). Moreover, assume that the Riemannian Hessian $\operatorname{Hess} f(A^*, B^*)\Big|_{\mu=0}$ is strictly positive definite as in Lemma 4.4.10 and that $0 < \delta < \lambda_{\min}(\operatorname{Hess}_A f(A^*, B^*))/4$. Then it follows:*

i. For all sufficiently large k, $\Delta^k(\tau^k) = \Delta_N^k$ is admissible in the backtracking dogleg search in step 3 of Algorithm 4.4.7; i.e. $A^{k+1} = P_{\bar{\mathcal{M}}}(A^k + \Delta_N^k)$ satisfies condition (4.4.1).

ii. The sequence $\{A^k\}$ converges q-linearly to A^* at rate κ_p; i.e.

$$\limsup_{k\to\infty} \frac{\|A^{k+1} - A^*\|}{\|A^k - A^*\|} \leq \kappa_p,$$

where $\kappa_p \in [0, 1)$ is a qualified constant in Lemma 4.4.10(iii) such that condition (4.4.12) holds.

iii. $\lim_{k\to\infty} \|P_{T_{\bar{\mathcal{M}}}(A^{k+1})}((1+\mu)A^{k+1} + B^k - Z)\| = 0$. Consequently, condition (4.4.3) is fulfilled with $\lim_{k\to\infty} \varepsilon_a^k = 0$.

Proof. By the continuity of the mapping $(A^k, B^k) \mapsto H^k$ and Lemma 4.4.10(i), we have $\lambda_{\min}(H^k) \geq \lambda_{\min}(\operatorname{Hess}_A f(A^*, B^*))/2 > 0$ for all sufficiently large k. Thus the backtracking dogleg search in step 3 of Algorithm 4.4.7 is initiated with $\tau^k = 2$, or $\Delta^k(\tau^k) = \Delta_N^k = -(H^k)^{-1}[g^k]$. Note that, due to Lemma 4.4.9(iii), both $\|g^k\|$ and $\|\Delta_N^k\|$ can be assumed to be sufficiently close to 0. Since $\Delta \in T_{\bar{\mathcal{M}}}(A) \mapsto P_{\bar{\mathcal{M}}}(A + \Delta) \in \bar{\mathcal{M}}$ is a second-order retraction on $\bar{\mathcal{M}}$ near A (see Example 18 in [AM12]), we have the following Taylor expansion:

$$\begin{aligned} f_A^k(P_{\bar{\mathcal{M}}}(A^k + \Delta_N^k)) &= f_A^k(A^k) + \langle g^k, \Delta_N^k\rangle + \frac{1}{2}\langle \Delta_N^k, H^k[\Delta_N^k]\rangle + o(\|\Delta_N^k\|^2) \\ &= f_A^k(A^k) - \frac{1}{2}\langle \Delta_N^k, H^k[\Delta_N^k]\rangle + o(\|\Delta_N^k\|^2), \quad \text{as } k \to \infty. \end{aligned}$$

Meanwhile, it follows from $A^{k+1} = P_{\bar{\mathcal{M}}}(A^k + \Delta_N^k) = A^k + \Delta_N^k + o(\|\Delta_N^k\|)$ that

$$\|A^{k+1} - A^k\|^2 = \|\Delta_N^k\|^2 + o(\|\Delta_N^k\|^2), \quad \text{as } k \to \infty.$$

Thus, altogether we have

$$\begin{aligned} f(A^k, B^k) - f(A^{k+1}, B^k) &= f_A^k(A^k) - f_A^k(A^{k+1}) = \frac{1}{2}\langle \Delta_N^k, H^k[\Delta_N^k]\rangle + o(\|\Delta_N^k\|^2) \\ &\geq \frac{\lambda_{\min}(\operatorname{Hess}_A f(A^*, B^*))}{4}\|\Delta_N^k\|^2 + o(\|\Delta_N^k\|^2) \\ &= \frac{\lambda_{\min}(\operatorname{Hess}_A f(A^*, B^*))}{4}\|A^{k+1} - A^k\|^2 + o(\|\Delta_N^k\|^2) \geq \delta\|A^{k+1} - A^k\|^2, \end{aligned}$$

i.e. $\Delta^k(\tau^k) = \Delta_N^k$ is admissible.

Now consider $A^{k+1} = \phi^A(A^k, B^k)$, where $\phi^A : \bar{\mathcal{M}} \times \mathcal{N} \to \bar{\mathcal{M}}$ is defined by the following system of equations:

$$\phi^A(A, B) = P_{\bar{\mathcal{M}}}(A + \Delta) =: \rho(A, \Delta), \tag{4.4.15}$$

$$\nabla_\Delta \mathrm{grad}_A f(A, B) = -\mathrm{grad}_A f(A, B). \tag{4.4.16}$$

If $(A^k, B^k) = (A^*, B^*)$, then we have $\mathrm{grad}_A f(A^*, B^*) = 0$. Moreover, since $\mathrm{Hess}_A f(A^*, B^*)$ is invertible, we have $\Delta = 0$ and thus $A^{k+1} = \rho(A^*, 0) = A^*$. Let us perturb ϕ^A at (A^*, B^*) with respect to the first argument along some $\Lambda_A \in T_{\bar{\mathcal{M}}}(A^*)$, which yields

$$D_A\phi^A(A^*, B^*)[\Lambda_A] = D_A\rho(A^*, 0)[\Lambda_A] + D_\Delta\rho(A^*, 0)[D_A\Delta(A^*, B^*)[\Lambda_A]].$$

Since $D_A\rho(A^*, 0)[\Lambda_A] = \Lambda_A$ and $D_\Delta\rho(A^*, 0)[\cdot] = \mathrm{id}_{T_{\bar{\mathcal{M}}}(A^*)}(\cdot)$ on $T_{\bar{\mathcal{M}}}(A^*)$, we have

$$D_A\phi^A(A^*, B^*)[\Lambda_A] = \Lambda_A + D_A\Delta(A^*, B^*)[\Lambda_A].$$

The function $\Delta(A, B)$ is implicitly defined through equation (4.4.16), and in particular $\Delta(A^*, B^*) = 0$.

Next we use a calculus approach to show the following identity:

$$D_A\mathrm{grad}_A f(A^*, B^*)[D_A\Delta(A^*, B^*)[\Lambda_A]] = -D_A\mathrm{grad}_A f(A^*, B^*)[\Lambda_A]. \tag{4.4.17}$$

Let Γ_A denote the matrix-form Christoffel symbols of $\bar{\mathcal{M}}$ around A (see, e.g., [EAS98]) such that Γ_A is symmetric, bilinear, and $\nabla_{\eta(A)}\xi(A) = D\xi(A)[\eta(A)] + \Gamma_A[\xi(A), \eta(A)] \in T_{\bar{\mathcal{M}}}(A)$ for any two smooth vector fields ξ and η on $\bar{\mathcal{M}}$. Then equation (4.4.16) can be rewritten as follows:

$$D_A\mathrm{grad}_A f(A, B)[\Delta(A, B)] + \Gamma_A[\mathrm{grad}_A f(A, B), \Delta(A, B)] = -\mathrm{grad}_A f(A, B).$$

By perturbing the above equation at (A^*, B^*) along $(\Lambda_A, 0) \in T_{\bar{\mathcal{M}}}(A^*) \times T_{\mathcal{N}}(B^*)$, we have

$$\begin{aligned}
&D_A^2\mathrm{grad}_A f(A^*, B^*)[\Delta(A^*, B^*), \Lambda_A] + D_A\mathrm{grad}_A f(A^*, B^*)[D_A\Delta(A^*, B^*)[\Lambda_A]] \\
&\quad + \Gamma_{A^*}[D_A\mathrm{grad}_A f(A^*, B^*)[\Lambda_A], \Delta(A^*, B^*)] + \Gamma_{A^*}[\mathrm{grad}_A f(A^*, B^*), D_A\Delta(A^*, B^*)[\Lambda_A]] \\
=& - D_A\mathrm{grad}_A f(A^*, B^*)[\Lambda_A].
\end{aligned}$$

Crossing out the vanishing terms, we obtain (4.4.17) as claimed. Note that $\mathrm{grad}_A f(A, B) = P_{T_{\bar{\mathcal{M}}}(A)}((1 + \mu)A + B - Z)$ and thus $D_A\mathrm{grad}_A f(A^*, B^*)[\cdot] = (1 + \mu)\mathrm{id}_{T_{\bar{\mathcal{M}}}(A^*)}(\cdot)$ on $T_{\bar{\mathcal{M}}}(A^*)$. Thus we have

$$D_A\phi^A(A^*, B^*) = 0.$$

Analogously, we perturb ϕ^A at (A^*, B^*) with respect to the second argument along $(0, \Lambda_B) \in T_{\bar{\mathcal{M}}}(A^*) \times T_{\mathcal{N}}(B^*)$. This leads to

$$D_B\phi^A(A^*, B^*)[\Lambda_B] = D_B\Delta(A^*, B^*)[\Lambda_B].$$

Again by the calculus approach, we derive

$$D_A \mathrm{grad}_A f(A^*, B^*)[D_B\Delta(A^*, B^*)[\Lambda_B]] = -D_B \mathrm{grad}_A f(A^*, B^*)[\Lambda_B].$$

Note that $D_B \mathrm{grad}_A f(A^*, B^*)[\Lambda_B] = P_{T_{\bar{\mathcal{M}}}(A^*)}(\Lambda_B)$. Then it follows that

$$D_B \phi^A(A^*, B^*)[\Lambda_B] = P_{T_{\bar{\mathcal{M}}}(A^*)}(\Lambda_B).$$

By the Taylor expansion of ϕ^A at (A^*, B^*), we have the following estimate

$$\begin{aligned} &\|A^{k+1} - A^*\| = \|\phi^A(A^k, B^k) - \phi^A(A^*, B^*)\| \\ &\le \|D_A\phi^A(A^*, B^*)(A^k - A^*)\| + \|D_B\phi^A(A^*, B^*)(B^k - B^*)\| + o(\|A^k - A^*\|) + o(\|B^k - B^*\|) \\ &= \|P_{T_{\bar{\mathcal{M}}}(A^*)}(B^k - B^*)\| + o(\|A^k - A^*\|) + o(\|B^k - B^*\|), \quad \text{as } k \to \infty. \end{aligned} \tag{4.4.18}$$

In order to obtain an estimate on $B^k - B^*$, consider the mapping $\phi^B(A, B) := P_{T_{\mathcal{N}}(B)}(A + B - Z)$. Let B^k be sufficiently close to B^* such that $\|B^k\|_0 = s$ and $B^k - B^* \in T_{\mathcal{N}}(B^*)$. Due to our assumption on the sequence $\{B^k\}$ and Theorem 4.4.4, we have $\phi^B(A^k, B^k) = \phi^B(A^*, B^*) = 0$. Moreover, the derivatives of ϕ^B are given by $D_A\phi^B(A^*, B^*) = P_{T_{\mathcal{N}}(B^*)}$ and $D_B\phi^B(A^*, B^*) = \mathrm{id}_{T_{\mathcal{N}}(B^*)}$. Thus the Taylor expansion of ϕ^B at (A^*, B^*) appears as

$$\begin{aligned} \phi^B(A^k, B^k) = \phi^B(A^*, B^*) + D_A\phi^B(A^*, B^*)(A^k - A^*) + D_B\phi^B(A^*, B^*)(B^k - B^*) \\ + o(\|A^k - A^*\|) + o(\|B^k - B^*\|), \quad \text{as } k \to \infty, \end{aligned}$$

which further implies that

$$B^k - B^* = -P_{T_{\mathcal{N}}(B^*)}(A^k - A^*) + o(\|A^k - A^*\|) + o(\|B^k - B^*\|), \quad \text{as } k \to \infty. \tag{4.4.19}$$

In particular, we have $\|B^k - B^*\| \le O(\|A^k - A^*\|)$ as $k \to \infty$.

By plugging (4.4.19) into (4.4.18), it follows from Lemma 4.4.10(iii) that

$$\begin{aligned} \|A^{k+1} - A^*\| &\le \|(P_{T_{\bar{\mathcal{M}}}(A^*)} \circ P_{T_{\mathcal{N}}(B^*)})(A^k - A^*)\| + o(\|A^k - A^*\|) \\ &\le \kappa_p\|A^k - A^*\| + o(\|A^k - A^*\|), \end{aligned}$$

for all sufficiently large k. This proves our claim (ii).

Finally, in view of the convergence of $\{(A^k, B^k)\}$ to (A^*, B^*) as well as Lemma 4.4.9(i), we conclude that $\lim_{k\to\infty} \|P_{T_{\bar{\mathcal{M}}}(A^{k+1})}((1 + \mu)A^{k+1} + B^k - Z)\| = 0$ and that condition (4.4.3) is fulfilled with $\lim_{k\to\infty} \varepsilon_a^k = 0$. □

We end this subsection by noting that the dependence of δ on (A^*, B^*) is certainly delicate. In our numerics, however, the choice of δ turned out to be rather unproblematic, even for $\mu = 0$ as in section 4.5. Concerning the complexity of the low-rank subproblem solver, note that the computation of Δ_N^k in step 2 of Algorithm 4.4.7 possibly requires solving the linear system involving H^k. Under the assumption on the positive definiteness of the Riemannian Hessian in Theorem 4.4.11, which imitates the second-order sufficient optimality condition in classical (unconstrained, Euclidean) optimization, each H^k-system solve, up to certain fixed tolerance of the error, can be carried out by the conjugate gradient method within a uniformly bounded number of iterations. Thus, the overall complexity for the low-rank subproblem solver at each iteration is no more than $O(mnr)$ flops. In addition, we remark that the constant κ_p in Lemma 4.4.10(iii), which in fact measures the angle between the tangent spaces $T_{\bar{\mathcal{M}}}(A^*)$ and $T_{\mathcal{N}}(B^*)$, is an intrinsic quantification of the local identifiability [CSPW11] at (A^*, B^*). Even though our alternating minimizer scheme solves its subproblems only inexactly, its asymptotical convergence rate (i.e. κ_p) is equally fast as that attained by the (exact) alternating projection method. When $0 \leq \kappa_p < 1$, (A^*, B^*) is a strict local minimizer, and $\{A^k\}$ converges to A^* q-linearly at rate κ_p, as shown in Theorem 4.4.11(ii). In case $\kappa_p = 0$, or equivalently $T_{\bar{\mathcal{M}}}(A^*)$ and $T_{\mathcal{N}}(B^*)$ are perpendicular to each other, the convergence of $\{A^k\}$ to A^* is even superlinear.

4.4.4 Alternating minimization scheme with trimming

The favorable performance of Algorithm 4.4.1 depends on a proper choice of r and s. If either r or s is too small, the constraint will rule out the desired solution. On the other hand, if either r or s is too large, the convergence property of Algorithm 4.4.1 is in danger due to the rank- or cardinality-deficiency at the desired solution. In this subsection we resolve this issue by incorporating a heuristic trimming procedure into the alternating minimization scheme which allows an adaptive tuning of r and s. The trimming of the matrix A^k is based on the k-means clustering algorithm [Seb84], and the trimming of the matrix B^k is based on a hard-thresholding.

In brief, we initialize the algorithm by some safe choices of r^1 and s^1 that are larger than the underlying r and s, respectively. As the iterates $A^k \in \bar{\mathcal{M}}(r^k)$ tend to settle, we partition the r^k largest singular values of A^k (in logarithmic scale) into two clusters by the k-means algorithm. If the gap between the means of the two clusters is larger than some prescribed threshold, then we set r^{k+1} to be the cardinality of the cluster of the larger mean, and replace the old A^k by its projection onto $\bar{\mathcal{M}}(r^{k+1})$. On the other hand, when the iterates $B^k \in \mathcal{N}(s^k)$ tend to stabilize along the sequence, we replace those entries of B^k, which are less than some threshold in absolute value, by 0 and set $s^{k+1} := \|B^k\|_0$. The detailed implementation of the alternating minimization scheme with trimming is specified in the following.

Algorithm 4.4.12 (Alternating minimization scheme with trimming)**.**
Choose $\delta > 0$, $\nu_a > 0$, $\nu_b > 0$, $\theta_a > 0$, $\theta_b > 0$. Initialize $r^1 \in \mathbb{N}$, $s^1 \in \mathbb{N}$, $A^0 \in \bar{\mathcal{M}}(r^1)$, $B^0 = P_{\mathcal{N}(s^1)}(Z - A^0)$. Set $k = 1$ and iterate:

1. Compute A^k as an approximate solution of the A-subproblem $\min_{A \in \bar{\mathcal{M}}(r^k)} \frac{1}{2}\|A + B^{k-1} - Z\|^2$ by Algorithm 4.4.7, which is represented in the compact SVD form $A^k = U^k \Sigma^k (V^k)^\top$.
2. Compute B^k as an approximate solution of the B-subproblem $\min_{B \in \mathcal{N}(s^k)} \frac{1}{2}\|A^k + B - Z\|^2$ by Algorithm 4.4.3.

3. If $\|A^k - A^{k-1}\|/\|A^{k-1}\| > \nu_a$, then set $r^{k+1} := r^k$; otherwise trim A^k as follows:

   (a) Partition the logarithms of the r^k largest singular values of A^k, namely $\{\log \sigma_j^k\}_{j=1}^{r^k}$, into two disjoint sets $\{\log \sigma_j^k\}_{j \in \mathcal{I}_1}$ and $\{\log \sigma_j^k\}_{j \in \mathcal{I}_2}$ by the k-means clustering algorithm (with $|\mathcal{I}_1| + |\mathcal{I}_2| = r^k$).

   (b) Evaluate the means of the two clusters; i.e. $m_1 := (\sum_{j \in \mathcal{I}_1} \log \sigma_j^k)/|\mathcal{I}_1|$ and $m_2 := (\sum_{j \in \mathcal{I}_2} \log \sigma_j^k)/|\mathcal{I}_2|$. Assume $m_1 \geq m_2$ without loss of generality.

   (c) If $m_1 - m_2 > \theta_a$, then set $r^{k+1} := |\mathcal{I}_1|$, $U^k := [\{U_j^k\}_{j \in \mathcal{I}_1}]$, $V^k := [\{V_j^k\}_{j \in \mathcal{I}_1}]$, $\Sigma^k := \mathrm{diag}(\{\sigma_j^k\}_{j \in \mathcal{I}_1})$, and $A^k := U^k \Sigma^k (V^k)^\top$.

4. If $\|B^k - B^{k-1}\|/\|B^{k-1}\| > \nu_b$, then set $s^{k+1} := s^k$; otherwise set $B_{ij}^k := 0$ whenever $|B_{ij}^k| < \theta_b$ and update $s^{k+1} := \|B^k\|_0$.

5. If a suitable stopping criterion is satisfied, then stop; otherwise increase k by 1 and return to step 1.

4.5 Numerical experiments

In this section, we study the numerical performance of Algorithm 4.4.12. The following parameters in the algorithm are fixed throughout the experiments: $\mu = 0$, $\delta = 0.1$, $\nu_a = \nu_b = 0.2$. Note that although it is favorable to consider $\mu > 0$ so as to guarantee the existence of a solution (see Theorem 4.3.1), we experience no troubles in our numerical experiments when choosing $\mu = 0$. Concerning the initialization, given any $A^0 \in \bar{\mathcal{M}}(r^1)$, we always take $B^0 = P_{\mathcal{N}(s^1)}(Z - A^0)$ accordingly. The inversion of the linear system (4.4.8) for computing the Newton step Δ_N^k is carried out by the conjugate gradient method with fixed residual tolerance 0.01. It turns out that this (approximate) Newton step in resolving the A-subproblem is so good that it is admissible, i.e. $\Delta^k(\tau^k) = \Delta_N^k$ fulfills condition (4.4.10), in almost every iteration. In addition, all partial SVDs are performed using the PROPACK routine `lansvd` [PRO], which should be distinguished from the (full) SVDs using the MATLAB routine `svd`.

The experiments were performed under MATLAB R2011b on a 2.66 GHz Intel Core Laptop with 4 GB RAM. All CPU-time reported in this section is measured in seconds.

4.5.1 Numerical behavior

We apply our algorithm to a test example of robust principal component pursuit. Let $m = n = 400$, $r = 0.05n$, $s = 0.05n^2$, and the observation matrix is generated by $Z = A_{true} + B_{true} + N$. The rank-$r$ matrix $A_{true} = L_{true} R_{true}^\top$ is generated by the product of two matrices $L_{true} \in \mathbb{R}^{m \times r}$ and $R_{true} \in \mathbb{R}^{n \times r}$, both of which have entries independently sampled from a normal distribution of mean 0 and standard deviation 1. The sparse matrix B_{true} has s nonzero entries, whose locations are randomly chosen and whose values are independently sampled from $\{\pm\sqrt{n}\}$ with uniform probability. The matrix N contains white Gaussian noise of mean 0 and standard deviation 0.001. In this example, we choose $\theta_a = \log 5$, $\theta_b = 0.2\sqrt{n}$.

Since our algorithm intends to find a local solution for the nonconvex minimization problem (4.3.1), it is important to check the quality of such a local solution as well as the dependence on the initial guess (A^0, B^0). In the following test, we consider two different choices for A^0, namely $A^0 = P_{\bar{\mathcal{M}}(r^1)}(Z)$ and A^0 being the projection of a random Gaussian matrix onto $\bar{\mathcal{M}}(r^1)$. Meanwhile, we also investigate the effectiveness of the trimming procedure for tuning r^k and s^k, provided that the true values of r and s are not available at the beginning. In this test, we allow r^1 and s^1 to be overestimations with respect to the true r and s up to 100%. The iterative algorithm is terminated once the relative error $\|A^k - A_{true}\|/\|A_{true}\|$ drops below 2×10^{-4}.

In Table 4.1, we report the corresponding relative error and the CPU-time. It is observed that the quality of the solutions produced by Algorithm 4.4.12, measured by the relative error, is robust to different initializations. Nonetheless, we remark that the efficiency of the algorithm is correlated to the choices of r^1, s^1, and A^0. As it can be expected, the initial guess $A^0 = P_{\bar{\mathcal{M}}(r^1)}(Z)$ is superior to a randomly chosen A^0 with respect to CPU-time, while choosing r^1 and s^1 closer to the underlying r and s yields faster convergence.

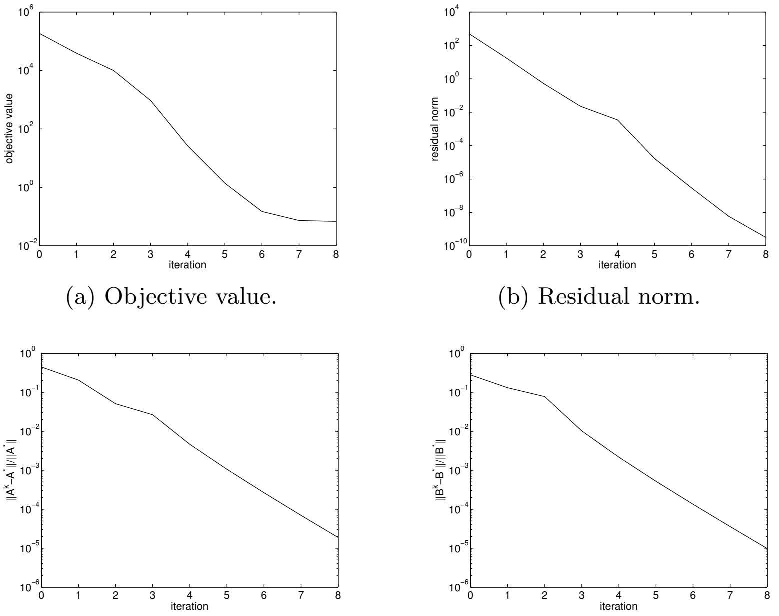
We further illustrate the numerical behavior of the algorithm, for instance, when $r^1 = 1.5r$, $s^1 = 1.5s$, and $A^0 = P_{\bar{\mathcal{M}}(r^1)}(Z)$. In Figure 4.1, we provide the semi-logarithmic plots of the objective value $f(A^k, B^k)$, the residual norm $\|\text{grad}_A f(A^k, B^k)\|$, and the convergence errors $\|A^k - A^*\|/\|A^*\|$ and $\|B^k - B^*\|/\|B^*\|$. The limit points A^* and B^* are precomputed with sufficiently high accuracy. It is observed from Figure 4.1(c) that, as is theoretically justified in Theorem 4.4.11, the sequence $\{A^k\}$ indeed exhibits a linear convergence, and the asymptotical convergence rate in this example is about 0.26.

r^1	s^1	$A^0 = P_{\bar{\mathcal{M}}(r^1)}(Z)$		random A^0	
		error	CPU	error	CPU
r	s	1.78e-4	1.32	1.19e-4	1.76
$1.25r$	$1.25s$	1.73e-4	1.74	1.67e-4	2.05
$1.5r$	$1.5s$	1.84e-4	1.77	1.35e-4	2.36
$1.75r$	$1.75s$	1.32e-4	2.03	1.01e-4	2.64
$2r$	$2s$	1.16e-4	2.17	1.83e-4	2.75

Table 4.1: Initialization study.

4.5.2 Comparison with an augmented Lagrangian method

A comprehensive comparison of numerical solvers on the (convex) nuclear-plus-ℓ^1-norm model for robust principal component pursuit can found on the webpage [LRM]. Among those solvers, the augmented Lagrangian method [CLMW11, LCWM09, TY11] seems to be the most efficient one in practice. Hence, in the following we compare the performances of our alternating minimization scheme and the augmented Lagrangian method with implementation-wise variations on both



(a) Objective value.

(b) Residual norm.

(c) Convergence of $\{A^k\}$.

(d) Convergence of $\{B^k\}$.

Figure 4.1: Convergence behavior.

methods. More specifically, we implement Algorithm 4.4.12 both with and without the global minimization trial step for the low-rank subproblem (i.e. step 0 of Algorithm 4.4.7), which are abbreviated by "AMS$^{\#}$" and "AMS" respectively. The implementation of the augmented Lagrangian method essentially follows Algorithm 1 in [CLMW11]. The major computational cost of this algorithm lies in an SVD in full dimension for performing a "singular value thresholding" at each iteration. As pointed out by [CCS10], it is possible to accelerate the singular value thresholding via partial SVD [PRO]. Different from the context in [CCS10], however, the target matrix (for SVD) in our matrix decomposition problem is dense and unstructured in general, and thus this acceleration strategy should only be utilized when the rank of the target matrix is predictably low. In our experiments, we implement the augmented Lagrangian method with full SVDs only (abbreviated by "fSVD-ALM"), and also its partial-SVD variant (abbreviated by "pSVD-ALM") where one switches from full SVD to partial SVD once the rank of the low-rank component A^k in the previous iteration drops below an empirical threshold equal to $0.2n$.

The test data is generated in the same way as described in the first paragraph of section 4.5.1, except for $N = 0$. Thus, the exact recovery of A_{true} and B_{true} is expected for all candidate methods, namely AMS, AMS$^{\#}$, fSVD-ALM, and pSVD-ALM. In this example, we choose $\theta_a = \log 5$, $\theta_b = 0.2\sqrt{n}$ in AMS and AMS$^{\#}$. Besides, we assume a moderate initial estimate (rather than the exact knowledge) of r and s such that $r^1 = 1.5r$, $s^1 = 1.5s$. For a fair comparison, we use the same initial guesses, i.e. $A^0 = P_{\bar{\mathcal{M}}(r^1)}(Z)$, $B^0 = P_{\mathcal{N}(s^1)}(Z - A^0)$, for all candidate methods. The experiments are performed with different combinations of n, r, and s.

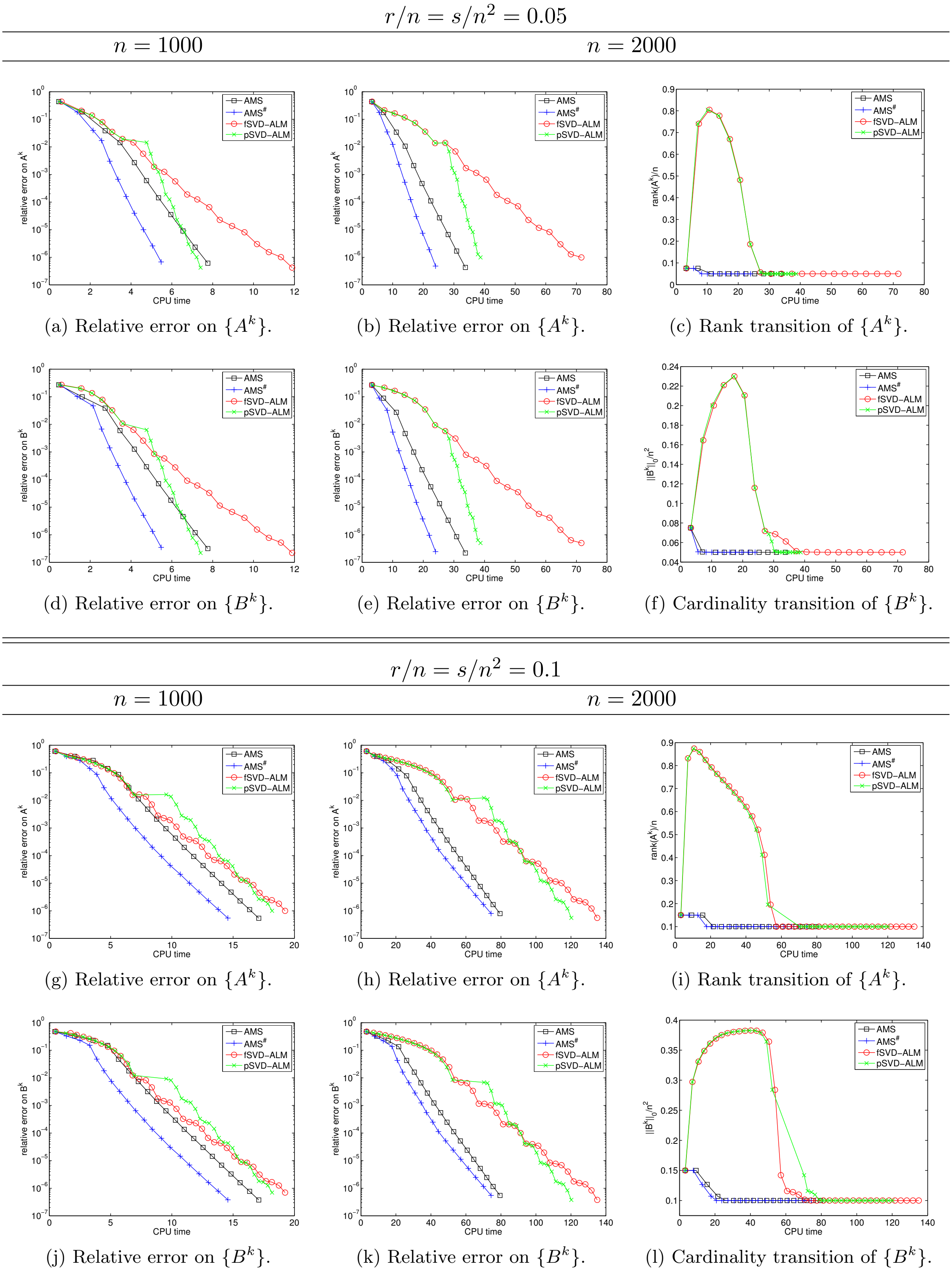
The corresponding comparisons among the four candidate methods with respect to relative errors, measured by $\|A^k - A_{true}\|/\|A_{true}\|$ and $\|B^k - B_{true}\|/\|B_{true}\|$, and CPU time are demonstrated in Figure 4.2. It is observed in the experiments that AMS$^\#$ always accepts the global minimizers from both the A- and B-subproblems, and essentially behaves like a heuristic alternating projection method (see the Appendix for a description) known to be locally linearly convergent. In this particular example, AMS$^\#$ works extremely well owing to a good initial guess so that the local convergence of the alternating projection method is immediately activated from the beginning. Nevertheless, the reader should be cautioned that in general such convergence behavior is not guaranteed for the alternating projection method with arbitrary initial guesses, and under such circumstances the global minimization trial steps are most likely wasteful. On the other hand, the plots on the relative errors in Figure 4.2 indicate that AMS, with guaranteed global convergence, has rather close performance to AMS$^\#$, especially for larger scales. Although partial SVDs typically improve the augmented Lagrangian method over the asymptotical convergence rate, expensive full SVDs are inevitable at early iterations; see the plots of the rank transitions of $\{A^k\}$ (when $n = 2000$) in the rightmost column of Figure 4.2. In comparison, AMS and AMS$^\#$ capture the rank of the low-rank component and the cardinality of the sparse component efficiently, thanks to the heuristic trimming procedure, and thus outperform fSVD-ALM and pSVD-ALM for large scales.

4.5.3 Application to background-foreground separation of surveillance video

We apply our algorithm to background-foreground separation of surveillance videos. Our first test video, which is taken from [LHGT04, CLMW11] and also publicly available [Sur], is a sequence of 200 frames taken in an airport. Each frame is a gray-level image of resolution 144 × 176, and is stacked as one column in the data matrix $Z \in \mathbb{R}^{25344\times 200}$; i.e. $m = 25344$, $n = 200$. Our goal is to extract from Z the static background (as the low-rank matrix A) and the moving foreground (as the sparse matrix B).

We implement the alternating minimization scheme (AMS) with $\theta_a = \log 10$, $\theta_b = 0.12$, $A^0 = P_{\bar{\mathcal{M}}(r^1)}(Z)$, $r^1 = 5$, and $s^1 \approx 0.1mn$, which is terminated once the residual norm $\|\text{grad}_A f(A^k, B^k)\|$ is reduced by a factor of 10^{-4}. It takes 39.4 seconds for AMS to converge, and the ultimate value of r^k is equal to 1 and $s^k \approx 0.0483mn$. The corresponding extractions for three selected frames are displayed in columns (b) and (c) in Figure 4.3. For comparison, we also perform the extraction using the augmented Lagrangian method (ALM). The implementation of ALM again follows [CLMW11], and we terminate the iterations once $\|Z - A^k - B^k\|/\|Z\| \leq 10^{-4}$. We note that only full SVDs are implemented in ALM, as partial SVDs do not lead to CPU gain in this problem. The results by ALM are shown in columns (d) and (e), and it takes 124.4 seconds for ALM to converge.

Our second example is a 400-frame sequence taken in a lobby with varying illumination [LHGT04, CLMW11, Sur]. Each frame is of resolution 128 × 160, and the data matrix Z is formulated as a 20480-by-400 matrix (i.e. $m = 20480$, $n = 400$). We run AMS with the same



$r/n = s/n^2 = 0.05$

$n = 1000$ $n = 2000$

(a) Relative error on $\{A^k\}$.

(b) Relative error on $\{A^k\}$.

(c) Rank transition of $\{A^k\}$.

(d) Relative error on $\{B^k\}$.

(e) Relative error on $\{B^k\}$.

(f) Cardinality transition of $\{B^k\}$.

$r/n = s/n^2 = 0.1$

$n = 1000$ $n = 2000$

(g) Relative error on $\{A^k\}$.

(h) Relative error on $\{A^k\}$.

(i) Rank transition of $\{A^k\}$.

(j) Relative error on $\{B^k\}$.

(k) Relative error on $\{B^k\}$.

(l) Cardinality transition of $\{B^k\}$.

Figure 4.2: Comparison with augmented Lagrangian method.

parameters as in the previous example except for $\theta_b = 0.06$, which is smaller than before, so that we allow more information in the sparse matrix. The algorithm converges after 69.19 seconds, and the ultimate value of r^k is equal to 2 and $s^k \approx 0.00413mn$. We also implement ALM using the same setting as before, for which it takes 193.5 seconds to converge. The separation results of both methods are displayed in Figure 4.4.

We conclude from the experiments that AMS performs well in background-foreground separation of surveillance videos, which is robust to the variation of illumination. In comparison with ALM, AMS typically eliminates the moving shadows in the backgrounds that occur in ALM, and provides sharper extractions of the moving foregrounds. Moreover, AMS has considerable advantage over ALM with respect to CPU-time.

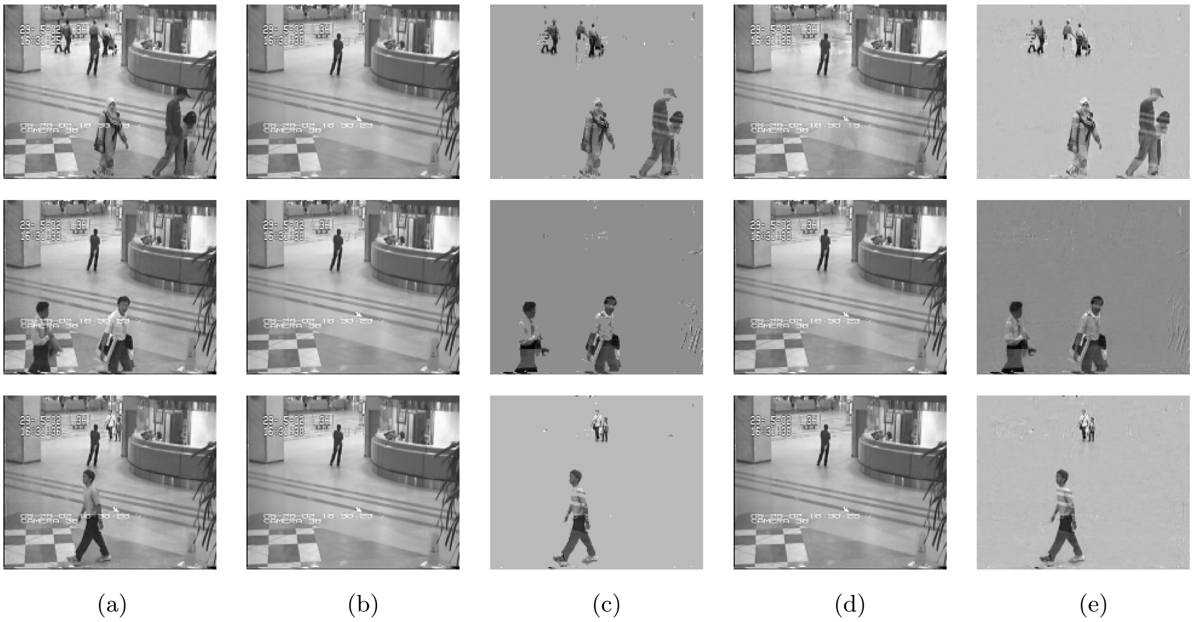


Figure 4.3: Background-foreground separation (airport): (a) original frames; (b) background via AMS; (c) foreground via AMS; (d) background via ALM; (e) foreground via AMS. The CPU-time consumed by AMS and ALM is 39.4 and 124.4 seconds, respectively.

4.6 Appendix on local convergence of an alternating projection method.

Here we consider a heuristic alternating projection method for the RPCP problem. This method, which can be interpreted as an exact alternating minimizer scheme for the optimization problem (1) with $\mu = 0$, can be shortly described as follows. Given $A^k \in \mathcal{M}$, one generates

$$\begin{cases} B^{k+1} := P_{\mathcal{N}}(Z - A^k), \\ A^{k+1} := P_{\mathcal{M}}(Z - B^{k+1}). \end{cases} \tag{4.6.1}$$

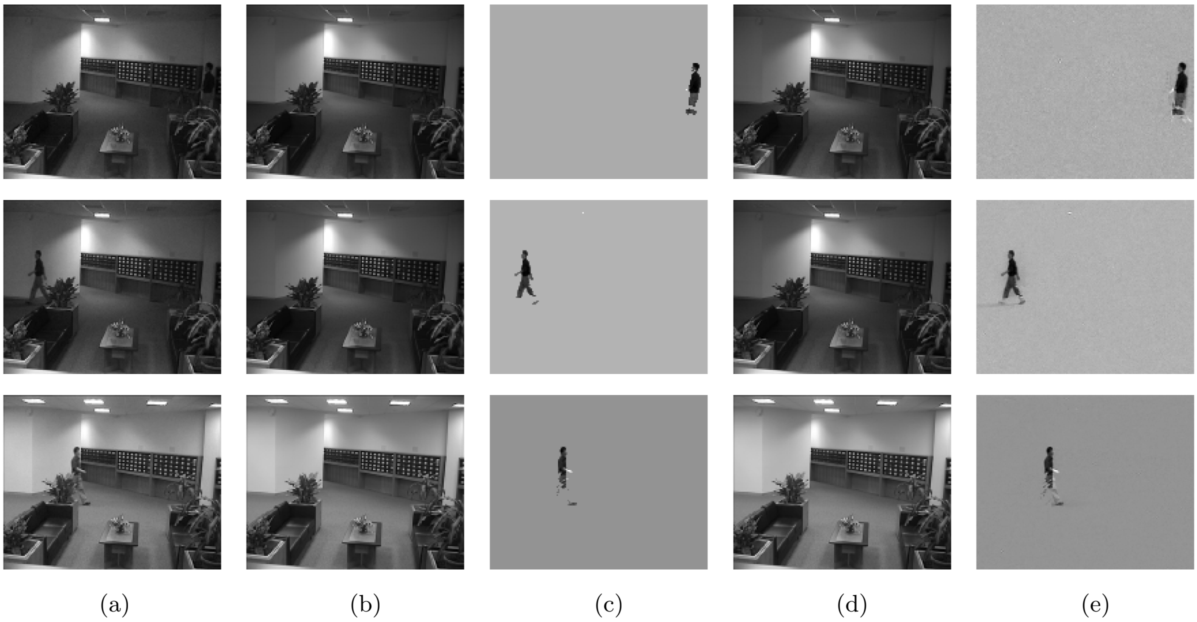


Figure 4.4: Background-foreground separation (lobby): (a) original frames; (b) background via AMS; (c) foreground via AMS; (d) background via ALM; (e) foreground via AMS. The CPU-time consumed by AMS and ALM is 69.19 and 193.5 seconds, respectively.

The name "alternating projection method" is termed, since the iterative procedure (on $\{A^k\}$) can be expressed as

$$A^{k+1} = \psi(A^k) := (P_{\mathcal{M}} \circ \iota \circ P_{\mathcal{N}} \circ \iota)(A^k), \tag{4.6.2}$$

with $\iota : A \mapsto Z - A$, and thus generalizes the classical alternating projection (where ι is the identity map) in, e.g, [LM08]. The following theorem asserts the local convergence of the alternating projection method. However, we note that the global convergence for this method is not guaranteed in general.

Theorem 4.6.1. *Given $A^0 \in \mathcal{M}$, let the sequence $\{(A^k, B^k)\}$ be iteratively generated by formula (4.6.1). Assume that (A^k, B^k) is sufficiently close to some (A^*, B^*) such that* $\operatorname{rank}(A^*) = r$, *$\|B^*\| = s$, $T_{\mathcal{M}}(A^*) \cap T_{\mathcal{N}}(B^*) = \{0\}$, and moreover*

$$\begin{cases} B^* := P_{\mathcal{N}}(Z - A^*), \\ A^* := P_{\mathcal{M}}(Z - B^*). \end{cases} \tag{4.6.3}$$

Then $\{(A^k, B^k)\}$ converges to (A^*, B^*) q-linearly at rate κ_p; i.e.

$$\limsup_{k\to\infty} \frac{\|(A^{k+1}, B^{k+1}) - (A^*, B^*)\|}{\|(A^k, B^k) - (A^*, B^*)\|} \le \kappa_p,$$

where $\kappa_p \in [0, 1)$ is a constant (same as in Lemma 4.4.10) such that

$$\|(P_{T_{\mathcal{M}}(A^*)} \circ P_{T_{\mathcal{N}}(B^*)})(\Delta)\| \le \kappa_p \|\Delta\|,$$

for all $\Delta \in \mathbb{R}^{m\times n}$.

Proof. We only prove the q-linear convergence on $\{A^k\}$, as the proof for $\{B^k\}$ is almost identical. Note that $\mathcal{M}$ and $\mathcal{N}$ are two smooth manifolds near A^* and B^*, respectively. For the existence of a qualified constant κ_p, we refer to Lemma 4.4.10(iii).

In the following, we perturb both equations in (4.6.3) with respect to A^* by an arbitrarily fixed $\Delta \in \mathbb{R}^{m\times n}$. The perturbation of the first equation gives

$$P_{\mathcal{N}}(Z - A^* - \Delta) = B^* + P_{\mathcal{N}}(Z - A^* - \Delta) - P_{\mathcal{N}}(Z - A^*) = B^* + P_{T_{\mathcal{N}}(B^*)}(-\Delta) + O(\|\Delta\|^2).$$

Since A^* is a fixed point of the map ψ in (4.6.2), the second equation in (4.6.3) can be written as $\psi(A^*) = P_{\mathcal{M}}(Z - B^*)$. Then we have

$$\begin{aligned}\psi(A^* + \Delta) &= P_{\mathcal{M}}(Z - P_{\mathcal{N}}(Z - A^* - \Delta)) = P_{\mathcal{M}}(Z - B^* - P_{T_{\mathcal{N}}(B^*)}(-\Delta) + O(\|\Delta\|^2)) \\ &= A^* + P_{\mathcal{M}}(Z - B^* - P_{T_{\mathcal{N}}(B^*)}(-\Delta) + O(\|\Delta\|^2)) - P_{\mathcal{M}}(Z - B^*) \\ &= A^* + (P_{T_{\mathcal{M}}(A^*)} \circ P_{T_{\mathcal{N}}(B^*)})(\Delta) + O(\|\Delta\|^2)).\end{aligned}$$

Thus, by considering $\Delta = A^k - A^*$ and passing $\Delta \to 0$, we conclude that

$$\limsup_{k\to\infty} \frac{\|A^{k+1} - A^*\|}{\|A^k - A^*\|} \leq \kappa_p.$$

□

Chapter 5

Conclusion and outlook

This thesis has investigated nonconvex and nonsmooth minimization methods in three different contexts, namely sparsity-promoting variational models with nonconvex priors, bilevel optimization with nonsmooth low-level problem, and optimization over Riemannian manifolds. Through the thesis, we conclude that such methods indeed yield advantages in a wide range of applications with respect to quality of the solutions or computational time. Nonconvex and nonsmooth minimizations certainly require more efforts, both analytically and numerically, in comparison with convex and/or smooth minimizations. Generally speaking, existence of solutions remains an open challenge in infinite dimensions. Characterization of optimality condition becomes challenging when the constraint set involves complex variational geometry. With careful design of solution algorithms, nonconvex and nonsmooth variational models can be numerically competitive in practice, provided that they are equipped with state-of-the-art techniques from optimization, analysis, and geometry.

To envisage our future research, we are interested in developing proper interpretation and utilization of nonconvex priors under the context of Bayesian statistics. The bilevel ansatz in the present thesis certainly admits extensions and applications to other imaging processing tasks. Interestingly, this includes a quasi-variational inequality approach to adaptively regularized image restoration. Besides, success of optimization algorithms over low-rank matrix manifolds inspires further challenge on optimization over low-rank tensor manifolds. This line of research is closely related to recent development in multilinear algebra.

Bibliography

[AA10] M. S. C. Almeida and L. B. Almeida. Blind and semi-blind deblurring of natural images. *IEEE Trans. Image Process.*, 19:36–52, 2010.

[ABM06] H. Attouch, G. Buttazzo, and G. Michaille. *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization.* SIAM, Philadelphia, 2006.

[ADM$^+$02] R. L. Adler, J.-P. Dedieu, J. Y. Margulies, M. Martens, and M. Shub. Newton's method on Riemannian manifolds and a geometric model for the human spine. *IMA J. Numer. Anal.*, 22:359–390, 2002.

[AK02] G. Aubert and P. Kornprobst. *Mathematical Problems in Image Processing.* Springer, New York, 2002.

[AM12] P.-A. Absil and J. Malick. Projection-like retractions on matrix manifolds. *SIAM J. Optim.*, 22:135–158, 2012.

[AMS08] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds.* Princeton University Press, New Jersey, 2008.

[AV94] R. Acar and C. R. Vogel. Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Problems*, 10:1217–1229, 1994.

[BA11] N. Boumal and P.-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *Advances in Neural Information Processing Systems*, volume 24, pages 406–414, 2011.

[BBC11] S. Becker, J. Bobin, and E. J. Candès. NESTA: A fast and accurate first-order method for sparse recovery. *SIAM J. Imaging Sci.*, 4:1–39, 2011.

[BDE09] A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev.*, 51:34–81, 2009.

[BJNP06] J. Bardsley, S. Jefferies, J. Nagy, and R. Plemmons. A computational method for the restoration of images with an unknown, spatially-varying blur. *Opt. Express*, 14:1767–1782, 2006.

[BMAS14] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *J. Mach. Learn. Res.*, 15:1455–1459, 2014.

[Boo03] W. M. Boothby. *An introduction to differentiable manifolds and Riemannian geometry*, volume Revised Second Edition. Academic Press, London, 2003.

[BS01] M. Burger and O. Scherzer. Regularization methods for blind deconvolution and blind source separation problems. *Math. Control Signals Systems*, 14:358–383, 2001.

[Car01] A. S. Carasso. Direct blind deconvolution. *SIAM J. Appl. Math.*, 61:1980–2007, 2001.

[Car02] A. S. Carasso. The APEX method in image sharpening and the use of low exponent Lévy stable laws. *SIAM J. Appl. Math.*, 63:593–618, 2002.

[Car06] A. S. Carasso. APEX blind deconvolution of color Hubble space telescope imagery and other astronomical data. *Optical Engineering*, 45:107004, 2006.

[Car09] A. S. Carasso. False characteristic functions and other pathologies in variational blind deconvolution: A method of recovery. *SIAM J. Appl. Math.*, 70:1097–1119, 2009.

[CCS10] J. F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, 20:1956–1982, 2010.

[CDS98] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20:33–61, 1998.

[CDS01] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43:129–159, 2001.

[CE07] P. Campisi and K. Egiazarian, editors. *Blind Image Deconvolution: Theory and Applications.* CRC press, Boca Raton, FL, 2007.

[CGM99] T. F. Chan, G. H. Golub, and P. Mulet. A nonlinear primal-dual method for total variation-based image restoration. *SIAM J. Sci. Comput.*, 20:1964–1977, 1999.

[CGT00] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust-Region Methods.* SIAM, Philadelphia, 2000.

[Cha07a] O. Chapelle. Training a support vector machine in the primal. *Neural Comput.*, 19:1155–1178, 2007.

[Cha07b] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Process. Lett.*, 14:707–710, 2007.

[Cha09] R. Chartrand. Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pages 262–265, 2009.

[Che12] X. Chen. Smoothing methods for nonsmooth, nonconvex minimization. *Math. Program., Ser. B*, 134:71–99, 2012.

[CK97] G. Chavent and K. Kunisch. Regularization of linear least squares problems by total bounded variation. *ESAIM Control Optim. Calc. Var.*, 2:359–376, 1997.

[CK12] C. Clason and K. Kunisch. A measure space approach to optimal source placement. *Comput. Optim. Appl.*, 53:155–171, 2012.

[Cla83] F. H. Clarke. *Optimization and Nonsmooth Analysis.* John Wiley & Sons, New York, 1983.

[CLMW11] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58:1–37, 2011.

[CM99] T. F. Chan and P. Mulet. On the convergence of the lagged diffusivity fixed point method in total variation image restoration. *SIAM J. Numer. Anal.*, 36:354–367, 1999.

[CNQ00] X. Chen, Z. Nashed, and L. Qi. Smoothing methods and semismooth methods for nondifferentiable operator equations. *SIAM J. Numer. Anal.*, 38:1200–1216, 2000.

[COS09] J.-F. Cai, S. Osher, and Z. Shen. Split Bregman methods and frame based image restoration. *Multiscale Model. Simul.*, 8:337–369, 2009.

[CP11] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40:120–145, 2011.

[CR09] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9:717–772, 2009.

[CS05] T. F. Chan and J. Shen. *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods.* SIAM, Philadelphia, 2005.

[CSPW11] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.*, 21:572–596, 2011.

[CT06] E. J. Candès and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory*, 52:5406–5425, 2006.

[CW98] T. F. Chan and C.-K. Wong. Total variation blind deconvolution. *IEEE Trans. Image Process.*, 7:370–375, 1998.

[CY08] R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3869–3872, 2008.

[CZ10] X. Chen and W. Zhou. Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization. *SIAM J. Imaging Sci.*, 3:765–790, 2010.

[DDFG10] I. Daubechies, R. DeVore, M. Fornasier, and C. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Comm. Pure Appl. Math.*, 63:1–38, 2010.

[DHN09] Y. Dong, M. Hintermüller, and M. Neri. An efficient primal-dual method for L^1TV image restoration. *SIAM J. Imaging Sci.*, 2:1168–1189, 2009.

[DHRC11a] Y. Dong, M. Hintermüller, and M. M. Rincon-Camacho. Automated regularization parameter selection in multi-scale total variation models for image restoration. *J. Math. Imaging Vis.*, 40:82–104, 2011.

[DHRC11b] Y. Dong, M. Hintermüller, and M. M. Rincon-Camacho. A multi-scale vectorial L^τ-TV framework for color image restoration. *Int. J. Comput. Vis.*, 92:296–307, 2011.

[DL92] D. L. Donoho and B. F. Logan. Signal recovery and the large sieve. *SIAM J. Appl. Math.*, 52:577–591, 1992.

[DlRS13] J. C. De los Reyes and C.-B. Schönlieb. Image denoising: Learning the noise model via nonsmooth PDE-constrained optimization. *Inverse Problems and Imaging*, 7:1183–1214, 2013.

[DM77] J. E. Dennis, Jr. and J. J. Moré. Quasi-Newton methods, motivation and theory. *SIAM Rev.*, 19:46–89, 1977.

[DM93] G. Dal Maso. *An Introduction to* Γ*-Convergence*. Birkhäuser, Boston, 1993.

[DS89] D. L. Donoho and P. B. Stark. Uncertainty principles and signal recovery. *SIAM J. Appl. Math.*, 49:906–931, 1989.

[DS96] J. E. Dennis, Jr. and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM, Philadelphia, 1996.

[EAS98] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20:303–353, 1998.

[ET99] I. Ekeland and R. Témam. *Convex Analysis and Variational Problems.* SIAM, Philadelphia, 1999.

[EY36] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

[FBP95] D. A. Fish, A. M. Brinicombe, and E. R. Pike. Blind deconvolution by means of the Richardson–Lucy algorithm. *J. Opt. Soc. Am. A*, 12:58–65, 1995.

[Fis97] A. Fischer. Solution of monotone complementarity problems with locally Lipschitzian functions. *Math. Program.*, 76:513–532, 1997.

[FL01] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[FLP98] M. Fukushima, Z.-Q. Luo, and J.-S. Pang. A globally convergent sequential quadratic programming algorithm for mathematical programs with linear complementarity constraints. *Comput. Optim. Appl.*, 10:5–34, 1998.

[FLRS06] R. Fletcher, S. Leyffer, D. Ralph, and S. Scholtes. Local convergence of SQP methods for mathematical programs with equilibrium constraints. *SIAM J. Optim.*, 17:259–286, 2006.

[GB82] E. M. Gafni and D. P. Bertsekas. Convergence of a gradient projection method. Laboratory for Information and Decision Systems Report LIDS-P-1201, Massachusetts Institute of Technology, 1982.

[Giu84] E. Giusti. *Minimal Surfaces and Functions of Bounded Variations*, volume 80 of *Monographs in Mathematics.* Birkhäuser, Boston, Basel, Stuttgart, 1984.

[GJY11] D. Ge, X. Jiang, and Y. Ye. A note on the complexity of L_p minimization. *Math. Program., Ser. B*, 129:285–299, 2011.

[GO09] T. Goldstein and S. Osher. The split Bregman method for L1-regularized problems. *SIAM J. Imaging Sci.*, 2:323–343, 2009.

[GR92] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14:367–383, 1992.

[HCSHJ12] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 57–60, 2012.

[HHM08] J. Huang, J. L. Horowitz, and S. Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.*, 36:587–613, 2008.

[HIK03] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13:865–888, 2003.

[HK04] M. Hintermüller and K. Kunisch. Total bounded variation regularization as a bilaterally constrained optimization problem. *SIAM J. Appl. Math.*, 64:1311–1333, 2004.

[HK09] M. Hintermüller and I. Kopacka. Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm. *SIAM J. Optim.*, 20:868–902, 2009.

[HMO05] L. He, A. Marquina, and S. J. Osher. Blind deconvolution using TV regularization and Bregman iteration. *International Journal of Imaging Systems and Technology*, 15:74–83, 2005.

[HR15] M. Hintermüller and C. N. Rautenberg. On the density of classes of closed, convex sets in Sobolev spaces arising from pointwise constraints on function values, the gradient or the divergence. *J. Math. Anal. Appl.*, 426:585–593, 2015.

[HS06] M. Hintermüller and G. Stadler. An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration. *SIAM J. Sci. Comput.*, 28:1–23, 2006.

[HS14] M. Hintermüller and T. Surowiec. A bundle-free implicit programming approach for a class of MPECs in function space. *preprint*, 2014.

[Hub64] P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 53:73–101, 1964.

[HW13] M. Hintermüller and T. Wu. Nonconvex TV^q-models in image restoration: Analysis and a trust-region regularization–based superlinearly convergent solver. *SIAM J. Imaging Sci.*, 6:1385–1415, 2013.

[HW14a] M. Hintermüller and T. Wu. A smoothing descent method for nonconvex TV^q-models. In *Lecture Notes in Computer Science*, volume 8293 of *Efficient Algorithms for Global Optimization Methods in Computer Vision*, pages 119–133. Springer, 2014.

[HW14b] M. Hintermüller and T. Wu. A superlinearly convergent R-regularized Newton scheme for variational models with concave sparsity-promoting priors. *Comput. Optim. Appl.*, 57:1–25, 2014.

[HW15a] M. Hintermüller and T. Wu. Bilevel optimization for calibrating point spread functions in blind deconvolution. *Inverse Problems and Imaging*, 9:1139–1169, 2015.

[HW15b] M. Hintermüller and T. Wu. Robust principal component pursuit via inexact alternating minimization on matrix manifolds. *J. Math. Imaging Vis.*, 51:361–377, 2015.

[IK99] K. Ito and K. Kunisch. An active set strategy based on the augmented Lagrangian formulation for image restoration. *ESAIM Math. Model. Num.*, 33:1–21, 1999.

[IK08] K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*. SIAM, Philadelphia, 2008.

[JCM12] K. Jia, T.-H. Chan, and Y. Ma. Robust and practical face recognition via structured sparsity. In *Proceedings of the 12th European Conference on Computer Vision*, volume 4, pages 331–344, 2012.

[JHSX11] H. Ji, S. Huang, Z. Shen, and Y. Xu. Robust video restoration by joint sparse and low rank matrix approximation. *SIAM J. Imaging Sci.*, 4:1122–1142, 2011.

[JR06] L. Justen and R. Ramlau. A non-iterative regularization approach to blind deconvolution. *Inverse Problems*, 22:771–800, 2006.

[Jus06] L. Justen. *Blind Deconvolution: Theory, Regularization and Applications.* PhD thesis, University of Bremen, 2006.

[KF00] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28:1356–1378, 2000.

[KH96a] D. Kundur and D. Hatzinakos. Blind image deconvolution. *IEEE Signal Process. Mag.*, 13:43–64, 1996.

[KH96b] D. Kundur and D. Hatzinakos. Blind image deconvolution revisited. *IEEE Signal Process. Mag.*, 13:61–63, 1996.

[KMO10a] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56:2980–2998, 2010.

[KMO10b] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, 2010.

[Knu97] D. Knuth. *The Art of Computer Programming*, volume 3: *Sorting and Searching.* Addison-Wesley, 3rd edition, 1997.

[KP13] K. Kunisch and T. Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM J. Imaging Sci.*, 6:938–983, 2013.

[KS01] A. C. Kak and M. Slaney. *Principles of Computerized Tomographic Imaging.* SIAM, Philadelphia, 2001.

[LCWM09] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. Technical Report UILU-ENG-09-2215, UIUC, 2009.

[Lev01] A. B. Levy. Solution sensitivity from general principles. *SIAM J. Control Optim.*, 40:1–38, 2001.

[LHGT04] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Trans. Image Process.*, 13:1459–1472, 2004.

[LM08] A. S. Lewis and J. Malick. Alternating projections on manifolds. *Math. Oper. Res.*, 33:216–234, 2008.

[LMa] LMaFit: Low-rank matrix fitting. `http://lmafit.blogs.rice.edu`.

[LPR96] Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints.* Cambridge University Press, 1996.

[LRM] Low-rank matrix recovery and completion via convex optimization. `http://perception.csl.illinois.edu/matrix-rank/sample_code.html`.

[MBS11] G. Meyer, S. Bonnabel, and R. Sepulchre. Regression on fixed-rank positive semidefinite matrices: A Riemannian approach. *J. Mach. Learn. Res.*, 12:593–625, 2011.

[Mor94] B. S. Mordukhovich. Generalized differential calculus for nonsmooth and set-valued mappings. *J. Math. Anal. Appl.*, 183:250–288, 1994.

[Mor06] B. S. Mordukhovich. *Variational analysis and generalized differentiation, I: Basic theory, II: Applications.* Springer, 2006.

[MZWM10] K. Min, Z. Zhang, J. Wright, and Y. Ma. Decomposing background topics from keywords by principal component pursuit. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 269–278, 2010.

[Nat95] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24:227–234, 1995.

[NC07] M. Nikolova and R. H. Chan. The equivalence of half-quadratic minimization and the gradient linearization iteration. *IEEE Trans. Image Process.*, 16:1623–1627, 2007.

[Nes05] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program., Ser. A*, 103:127–152, 2005.

[Nik99] M. Nikolova. Markovian reconstruction using a GNC approach. *IEEE Trans. Image Process.*, 8:1204–1220, 1999.

[Nik02] M. Nikolova. Minimizers of cost-functions involving nonsmooth data-fidelity terms. application to the processing of outliers. *SIAM J. Numer. Anal.*, 40:965–994, 2002.

[Nik05] M. Nikolova. Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares. *Multiscale Model. Simul.*, 4:960–991, 2005.

[NNT10] M. Nikolova, M. K. Ng, and C.-P. Tam. Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction. *IEEE Trans. Image Process.*, 19:3073–3088, 2010.

[NNZC08] M. Nikolova, M. K. Ng, S. Zhang, and W.-K. Ching. Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization. *SIAM J. Imaging Sci.*, 1:2–25, 2008.

[NS12] T. T. Ngo and Y. Saad. Scaled gradients on Grassmann manifolds for matrix completion. In *Advances in Neural Information Processing Systems*, volume 25, pages 1421–1429, 2012.

[NW06] J. Nocedal and S. Wright. *Numerical Optimization.* Springer, New York, 2nd edition, 2006.

[OKZ98] J. Outrata, M. Kocvara, and J. Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[O'n83] B. O'neill. *Semi-Riemannian Geometry: with Applications to Relativity.* Academic Press, London, 1983.

[Out00] J. V. Outrata. A generalized mathematical program with equilibrium constraints. *SIAM J. Control Optim.*, 38:1623–1638, 2000.

[PRO] PROPACK—Software for large and sparse SVD calculations. `http://sun.stanford.edu/~rmunk/PROPACK/`.

[QS93] L. Qi and J. Sun. A nonsmooth version of Newton's method. *Math. Program.*, 58:353–367, 1993.

[Rob80] S. M. Robinson. Strongly regular generalized equations. *Math. Oper. Res.*, 5:43–62, 1980.

[Rob87] S. M. Robinson. Local structure of feasible sets in nonlinear programming, Part III: Stability and sensitivity. *Math. Programming Stud.*, 30:45–66, 1987.

[ROF92] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[RW98] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, New York, 1998.

[Sch01] S. Scholtes. Convergence properties of a regularization scheme for mathematical programs with complementarity constraints. *SIAM J. Optim.*, 11:918–936, 2001.

[SE10] L. Simonsson and L. Eldén. Grassmann algorithms for low rank approximation of matrices with missing values. *BIT Numer. Math.*, 50:173–191, 2010.

[Seb84] G. A. F. Seber. *Multivariate Observations*. John Wiley & Sons, Hoboken, NJ, 1984.

[Sha05] A. Shapiro. Sensitivity analysis of parameterized variational inequalities. *Math. Oper. Res.*, 30:109–126, 2005.

[Smi93] S. T. Smith. *Geometric Optimization Methods for Adaptive Filtering*. PhD thesis, Harvard University, 1993.

[SS00] H. Scheel and S. Scholtes. Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity. *Math. Oper. Res.*, 25:1–22, 2000.

[Sta09] G. Stadler. Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices. *Comput. Optim. Appl.*, 44:159–181, 2009.

[Sur] Statistical modeling of complex background for foreground object detection. `http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html`.

[TB97] L. N. Trefethen and D. Bau, III. *Numerical Linear Algebra*. SIAM, Philadelphia, 1997.

[TW10] J. A. Tropp and S. J. Wright. Computational methods for sparse solution of linear inverse problems. In *Proceedings of the IEEE*, volume 98, pages 948–958, 2010.

[TY11] M. Tao and X. Yuan. Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM J. Optim.*, 21:57–81, 2011.

[Ulb03] M. Ulbrich. Semismooth newton methods for operator equations in function spaces. *SIAM J. Optim.*, 13:805–841, 2003.

[Van13] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM J. Optim.*, 23:1214–1236, 2013.

[VO96] C. R. Vogel and M. E. Oman. Iterative methods for total variation denoising. *SIAM J. Sci. Comput.*, 17:227–238, 1996.

[Vog02] C. R. Vogel. *Computational Methods for Inverse Problems.* SIAM, Philadelphia, 2002.

[WMC$^+$00] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs. In *Advances in Neural Information Processing Systems*, volume 13, pages 668–674, 2000.

[Wri97] S. J. Wright. *Primal-Dual Interior-Point Methods.* SIAM, Philadelphia, 1997.

[WYZ12] Z. Wen, W. Yin, and Y. Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Math. Prog. Comp.*, 4:333–361, 2012.

[Ye05] J. J. Ye. Necessary and sufficient optimality conditions for mathematical programs with equilibrium constraints. *J. Math. Anal. Appl.*, 307:350–369, 2005.

[YK96] Y.-L. You and M. Kaveh. A regularization approach to joint blur identification and image restoration. *IEEE Trans. Image Process.*, 5:416–428, 1996.

[YZZ97] J. J. Ye, D. L. Zhu, and Q. J. Zhu. Exact penalization and necessary optimality conditions for generalized bilevel programming problems. *SIAM J. Optim.*, 7:481–507, 1997.

[ZGLM12] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma. TILT: Transform invariant low-rank textures. *Int. J. Comput. Vis.*, 99:1–24, 2012.