Inaugural-Dissertation

submitted to the

Combined Faculty for the
Natural Sciences and Mathematics

of

Heidelberg University, Germany

for the degree of
Doctor of Natural Sciences

put forward by
M.Sc. Jan Plier
born in Crailsheim

Date of oral examination: .....

Theoretical and Numerical Approaches to Co-/Sparse Recovery in Discrete Tomography

Advisor: Prof. Dr. Stefania Petra

Zusammenfassung

Im Rahmen der vorliegenden Dissertation werden theoretische und numerische Ergebnisse hergeleitet, welche die exakte Rekonstruktion von stückweise konstanten Bildern aus unzureichenden Messungen in der diskreten Tomographie garantieren. Dies ist häufig der Fall bei der zerstörungsfreien Qualitätsprüfung von Industriebauteilen, die aus wenigen homogenen Materialien bestehen und bei denen aufgrund von schnellen Prüfzeiten keine vollständige Messung möglich ist. Die geringe Anzahl an Messungen führt schließlich zu einen unterbestimmten linearen Gleichungssystem. Es wird ein beschränkter Lösungsraum betrachtet, in welchem die Lösungen (a) dünnbesetzte (sparse) Bildgradienten aufweisen und (b) beschränkte Pixelwerte besitzen.

Auf der Grundlage der „Compressed Sensing“-Theorie wird eine untere Schranke für die Anzahl der Messungen bestimmt, die erforderlich sind, um unter den gegebenen Bedingungen, mittels konvexer Optimierung ein Bild eindeutig zu rekonstruieren. Darüber hinaus wird auch für den nicht-konvexen Fall eine untere Schranke bestimmt. Deren Besonderheit darin besteht, dass sie auf der Anzahl der Zusammenhangskomponenten beruht.

Auf diesen Ergebnissen aufbauend werden verschiedene Optimierungsmodelle untersucht, welche Lösungen mit sparsen Bildgradienten liefern oder den Bereich der Pixelwerte einschränken. In diesem Zusammenhang wird eine neuartige konvexe Relaxierung vorgestellt, die nachweislich genauer ist als alle bestehenden Ansätze. Hierbei wird davon ausgegangen, dass das Lösungsbild einen sparsen Gradienten aufweist und ganzzahlig ist. Da die Anzahl der Zusammenhangskomponenten in einem Bild für dessen eindeutige Rekonstruktion entscheidend ist, wird ein ganzzahliges Programm entwickelt, welches die maximale Anzahl der Zusammenhangskomponenten im rekonstruierten Bild beschränkt. Ferner wird beim Lösen konvexer Modelle der Bildbereich als Mannigfaltigkeit aufgefasst. Basierend auf Ergebnissen aus der Differentialgeometrie und der Optimierung auf Mannigfaltigkeiten wird ein Optimierungsverfahren erster Ordnung, welches mehrere Ebenen nutzt, hergeleitet. Der entwickelte Mehrebenen-Algorithmus weist eine schnelle Konvergenz auf und ermöglicht die Rekonstruktion von hochauflösenden Bildern.

Abstract

We investigate theoretical and numerical results that guarantee the exact reconstruction of piecewise constant images from insufficient projections in Discrete Tomography. This is often the case in non-destructive quality inspection of industrial objects, made of few homogeneous materials, where fast scanning times do not allow for full sampling. As a consequence, this low number of projections presents us with an underdetermined linear system of equations. We restrict the solution space by requiring that solutions (a) must possess a sparse image gradient, and (b) have constrained pixel values.

To that end, we develop an lower bound, using compressed sensing theory, on the number of measurements required to uniquely recover, by convex programming, an image in our constrained setting. We also develop a second bound, in the non-convex setting, whose novelty is to use the number of connected components when bounding the number of linear measurements for unique reconstruction.

Having established theoretical lower bounds on the number of required measurements, we then examine several optimization models that enforce sparse gradients or restrict the image domain. We provide a novel convex relaxation that is provably tighter than existing models, assuming the target image to be gradient sparse and integer-valued. Given that the number of connected components in an image is critical for unique reconstruction, we provide an integer program model that restricts the maximum number of connected components in the reconstructed image. When solving the convex models, we view the image domain as a manifold and use tools from differential geometry and optimization on manifolds to develop a first-order multilevel optimization algorithm. The developed multilevel algorithm exhibits fast convergence and enables us to recover images of higher resolution.

Acknowledgment

I want to thank various people who supported me during my Ph.D.

First of all, I thank my advisor, Prof. Stefania Petra, at the Faculty of Mathematics and Computer Science at Heidelberg University for her guidance, continuous support, as well as for sharing her comprehensive knowledge.

During my Ph.D., I stayed at the University of Birmingham for one month. I thank Prof. Michal Kočvara for our fruitful discussions, which have given me valuable insights into multilevel optimization.

Thanks for the support and funding of the German Research Foundation (DFG). I was a member of the Research Training Group (RTG) 1653 "Spatio/Temporal Probabilistic Graphical Models and Applications in Image Analysis", which I acknowledge gratefully.

Next, I want to thank Dr. Paul Swoboda, who contributed a lot at the beginning of my Ph.D. I also thank Evelyn and Barbara for all the assistance in administrative matters and the kind and cheerful chats in between. I want to mention my former and present colleagues Lukas, Artjom, Matthias, Francesco, Alexander, Fabrizio, Ruben, Dimitrij, Jörg, Fabian, Ecaterina, Freddie, and Andreas. They contributed explicitly by discussions and sharing their knowledge, or implicitly by having cake breaks. In particular, Fabrizio, who is the source of my knowledge about differential geometry. Special thanks to Artjom, Francesco, and Lukas for the weekly card game. Thanks to my sister in law, Sarah, and Lukas for proofreading.

I thank my wife, Christina, for her loving support and compassion throughout my Ph.D. Also, for encouraging me to continue when things got too frustrating.

Contents

List of Figures

List of Tables

List of Algorithms

List of Symbols

Sets

$\mathbb{R}$	real numbers
$\mathbb{Z}$	integer numbers
$\mathbb{N}$	natural numbers
$\overline{\mathbb{R}}$	$:= \mathbb{R} \cup \{\infty\}$, extended real numbers
$[n]$	$:= \{0, \ldots, n-1\}$ for $n \in \mathbb{N}$
$[a:b]$	$:= \{a, a+1, a+2, \ldots, b\}$ for $a, b \in \mathbb{Z}$ with $a < b$

Norms

$\|x\|_1$	$:= \sum_{i \in [n]} \|x_i\|$ for $x \in \mathbb{R}^n$
$\|x\|_2$	$:= \sqrt{\sum_{i \in [n]} x_i^2}$ for $x \in \mathbb{R}^n$
$\|x\|_\infty$	$:= \max_{i \in [n]} \|x_i\|$ for $x \in \mathbb{R}^n$

Analysis

$f\|_D$	$: D \to \overline{\mathbb{R}}$, restriction of function $f : \mathbb{R}^n \to \mathbb{R}$ to $D \subseteq \mathbb{R}^n$
$\operatorname{dom} f$	$:= \{x \in \mathbb{R}^n \mid f(x) < \infty\}$, where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$
$\nabla f(x)$	gradient of f at x
∇x	discrete gradient of the vector $x \in \mathbb{R}^n$ (see (5.13))

Linear Algebra

$\mathbb{1}$	$:= (1, \ldots, 1)^\top$
I	identity matrix
$\operatorname{aff}(C)$	affine hull of the set $C \subseteq \mathbb{R}^n$
$\mathcal{R}(A)$	$:= \{Ax \mid x \in \mathbb{R}^n\}$ is the range of the matrix $A \in \mathbb{R}^{m \times n}$
$\mathcal{N}(A)$	$:= \{x \in \mathbb{R}^n \mid Ax = 0\}$ is the nullspace of the matrix $A \in \mathbb{R}^{m \times n}$
$\operatorname{rank}(A)$	number of linear independent rows of the matrix $\mathbb{R}^{m \times n}$

Set Operations

$\overline{C}$	closure of the set $C \subseteq \mathbb{R}^n$
2^D	$:= \{C \mid C \subseteq D\}$ is the power set of the set D
$C^\infty(D)$	set of all smooth functions $f : D \to D$
$\operatorname{int}(C)$	interior of the set $C \subseteq \mathbb{R}^n$
$\operatorname{relint}(C)$	relative interior of the set $C \subseteq \mathbb{R}^n$

Convex Analysis

C°	$:= \{x \in \mathbb{R}^n \mid \langle x, y\rangle \leq 0, \forall y \in C\}$, polar cone of a convex set $C \subseteq \mathbb{R}^n$
$\operatorname{cone}(C)$	$:= \{\lambda x \mid \lambda \in \mathbb{R}_{\geq 0}, x \in C\}$ is the conic hull of a convex set $C \subseteq \mathbb{R}^n$
$\mathcal{D}_f(x)$	$:= \operatorname{cone}(\{z - x \colon f(z) \leq f(x)\})$ is the descent cone (see Def. 3.15)
$N(\overline{x}, C)$	$:= \{g \in \mathbb{R}^n \mid \langle g, y - \overline{x}\rangle \leq 0, \forall y \in C\}$ is the normal cone of C at x
$\partial f(x)$	set of subgradients at x of f (see Def. 2.3)
$\delta_C(x)$	indicator function of the set C (see (2.9))
$\operatorname{dist}(x, C)$	$:= \min_{y \in C} \|x - y\|_2$ of a closed convex set $C \subseteq \mathbb{R}^n$
$\pi_C(x)$	$:= \arg\min_{y \in C} \|x - y\|_2$, x projected onto the convex set $C \subseteq \mathbb{R}^n$

Probability Theory

$X \sim \mathcal{N}(0, 1)$	X is a normal distributed random variable
$\mathbb{E}[X]$	expected value of the random variable X

Compressed Sensing

$A_{S,\bullet}$	chooses rows from the matrix $A \in \mathbb{R}^{m \times n}$ by the index set $S \subseteq [m]$
$A_{\bullet,S}$	chooses columns from the matrix $A \in \mathbb{R}^{m \times n}$ by the index set $S \subseteq [n]$
x_S	chooses the entries from $x \in \mathbb{R}^n$ by the index set $S \subseteq [n]$
$\operatorname{sign}(x)$	computes componentwise $\frac{x_i}{\lvert x_i\rvert}$ if $x_i \neq 0$ and 0 otherwise for $x \in \mathbb{R}^n$
$\delta(C)$	statistical dimension of the cone $C \subseteq \mathbb{R}^n$ (see Def. 3.12)
$\operatorname{supp}(x)$	$:= \{i \in [n] \mid x_i \neq 0\}$ is the support of the vector $x \in \mathbb{R}^n$
$\operatorname{cosupp}(x)$	$:= \{i \in [n] \mid x_i = 0\}$ is the cosupport of the vector $x \in \mathbb{R}^n$
$\|x\|_0$	$:= \lvert\operatorname{supp}(x)\rvert$ for $x \in \mathbb{R}^n$

Differential Geometry

$\mathcal{M}, \mathcal{N}$	smooth manifolds (see Def. 2.9)
$T_p\mathcal{M}$	tangent space of $\mathcal{M}$ at $p \in \mathcal{M}$ (see (2.15))
$d_x f$	differential of $f : \mathcal{M} \to \mathcal{N}$ at $x \in \mathcal{M}$ (see (2.16))
$\nabla_{\mathcal{M}} f(x)$	Riemannian gradient of f at x (see (2.20))
$\mathfrak{R}_x$	$: T_p\mathcal{M} \to \mathcal{M}$, retraction on $\mathcal{M}$ (see Def. 2.11)
$\operatorname{id}_{\mathcal{M}}$	the identity map, i.e., $\operatorname{id}_{\mathcal{M}} : \mathcal{M} \to \mathcal{M}$

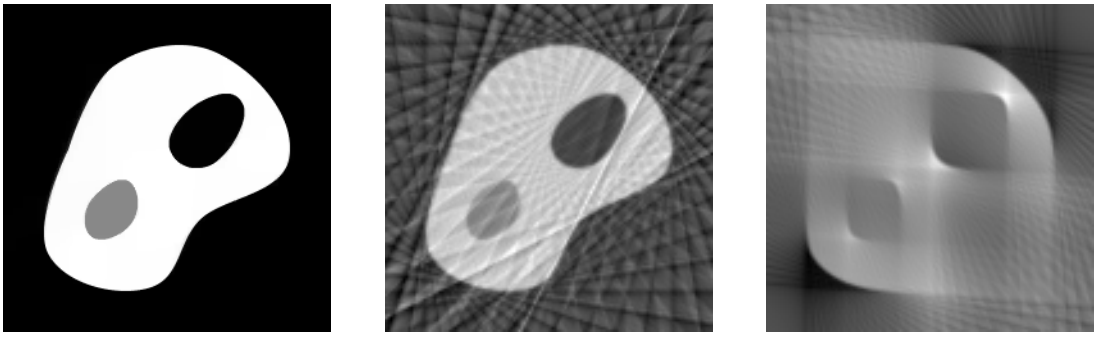
CHAPTER 1

Introduction

Computerized Tomography [Her09] (CT) is a widely used non-invasive method for creating cross-section images of an object. Initially developed for medical applications, CT was soon used for industrial applications as well. For instance, in quality inspection, one chooses a sample of produced parts, which is then destroyed to check for cracks or other defects. This results in financial loss when the selected sample was non-defective. By implementing CT for non-destructive quality inspection in a production line, every produced part is checked, and only defect parts are discarded.

CT uses *X-rays*, which are transmitted through an object and partially absorbed by material or tissue. For each X-ray the intensity that was not absorbed is measured by a detector. Given these intensities, we reconstruct the attenuation at each spatial position. The classical approach to tomographic reconstruction is the filtered back-projection (FBP). For high-resolution images, FBP requires X-ray measurements at numerous angles in order to comply to classical sampling theorems [Nat01]. In industrial applications, this requirement cannot be accommodated due to time or physical restrictions. Typically, in industrial applications, one affords only few projections rendering FBP impractical. Figure 1.1 illustrates the poor reconstruction quality when FBP is applied in limited-data or limited-angle scenarios.

Discrete tomography (DT) [HK99; HK07] is the mathematical field dealing with image reconstruction from CT in the case of limited-angle or limited-data, and is an active area of research to this day [BS11; BDP17; Den14b; FQ13; Gou13; Kap15; KSP17; LB16; SJP11; WSH03; Zis16]. It considers the reconstruction of an $d \times d$ image that represents the desired discrete cross-section image. The $n := d^2$ coefficients of the pixels are given by a vector $x \in \mathbb{R}^n$, where each x_j with $j \in [n]$ corresponds to the attenuation in the region described by pixel j. Then, we discretize the X-rays according to a $d \times d$ pixel grid, and the coefficients $A_{ij} \in \mathbb{R}$ provide an incidence geometry defining how pixel j contributes to the i-th X-ray. Thus, let $\overline{x} \in \mathbb{R}^n$ be the



(a) original (b) limited-data (c) limited-angle

Figure 1.1: Comparison of the reconstruction quality in *Computer Tomography* using filtered backprojection when the number of projections is small (b) or the range of angles are limited (c).

unknown vectorized image. Then the image reconstruction problem in DT corresponds to solving the linear system

$$Ax = A\overline{x} =: b, \tag{1.1}$$

where $A \in \mathbb{R}^{m \times n}$ models m X-rays, each row one X-ray.

In the situation of limited-angle or limited-data, the linear system (1.1) is highly under-determined, i.e., $m \ll n$. Hence, the inverse problem posed by (1.1) is ill-posed as it does not admit a unique solution. To remedy this situation, we need to include prior knowledge about the image to be reconstructed.

In view of our motivating industrial application, i.e., non-destructive quality inspection, the objects consist of few materials and the resulting cross-section image exhibits large homogeneous regions. Based on these observations, we assume that the images in question

(a) must possess a sparse image gradient, and
(b) have constrained pixel values.

The goal of this thesis is to investigate theoretical and numerical results that guarantee the exact reconstruction subject to (a) and (b). On the theoretical side, we provide lower bounds on the number of measurements that concern unique recovery by convex or non-convex optimization. Having established theoretical lower bounds on the number of required measurements, we then examine several optimization models that enforce sparse gradients or restrict the image domain.

In particular, we exploit results from the field of *Compressed Sensing* (CS) [FR13]. It provides recovery guarantees related to the sparsity of a signal and the number of linear measurements. However, most results hold for randomized measurements, unlike tomography that uses deterministic sensors. Existing results for deterministic measurements, e.g., *expander graphs* [FR13], only apply for specialized geometries [PS14; PSS13]. Moreover, these results do not extend to gradient sparse images. Though

the results from [NW13a; NW13b] are concerned with the recovery of gradient sparse images, the work [Den14a] shows that they lead to overly pessimistic guarantees for DT. Hence, classical existing recovery guarantees from CS do not apply. Nevertheless, CS implies that the sparsity of an image or the image gradient is a crucial property to guarantee unique reconstruction. Therefore, we use tools from CS to provide lower bounds on the necessary number of linear measurements considering sparse image gradients and constrained pixel values. By using phase transitions [Ame14], we empirically show that our theoretical lower bounds are close to the empirical success rates of DT reconstructions.

Additionally, we provide lower bounds for unique reconstruction when using non-convex optimization. Here, we use the theory of *union of subspaces* [BD09; LD08]. As a result, we show that the number of connected components in an image is a crucial parameter in predicting the needed number of linear measurements, which was not considered before in the literature.

Motivated by the theoretical lower bounds, we then examine several optimization models that enforce sparse gradients or restrict the image domain. Our results are reinforced by the authors in [Kei17] who showed that the lower bound is further reduced if the pixel values are restricted to a finite set. Hence, we provide a novel convex relaxation that is provably tighter than existing models, assuming the target image to be gradient sparse and integer-valued. Given that the number of connected components in an image is critical for unique reconstruction, we provide an integer program model that restricts the maximum number of connected components in the reconstructed image. Moreover, we derive inequalities to speed-up the convergence of the branch and bound algorithm used to solve integer programs. When solving the convex models, we view the image domain as a manifold and use tools from differential geometry and optimization on manifolds to develop a first-order multilevel optimization algorithm. The developed multilevel algorithm exhibits fast convergence and allows us to consider images of higher resolution.

1.1 Outline

Each chapter of this thesis is mainly self-contained and covers a different aspect of image reconstruction in *Discrete Tomography*.

Chapter 2 serves as an introduction to recall definitions needed throughout this thesis and provides further literature. Proofs are generally omitted but are given in the cited literature.

In Chapter 3, we use phase transitions to examine the success rate for unique recovery in discrete tomography by convex programming. We derive a dual certificate for generalized box-constrained convex optimization models, which is sufficient for proving uniqueness for a given solution. Since the existing theoretical lower bounds for unique recovery by convex programming do not apply to box-constraints, or non-negativity constraints, we prove an extension to cover these cases. Finally, we provide empirical

evidence that the proven lower bound is close to the phase transition of tomographic reconstruction.

In Chapter 4, motivated by the lower bounds, we develop a novel convex relaxation to reconstruct non-binary integer-valued images from tomographic projections. Further, we provide an approach to show that our convex relaxation is theoretical and empirical tighter than existing ones.

In Chapter 5, we consider sparse reconstruction by a non-convex optimization model. Using the theory of *union of subspaces* [LD08; BD09], we devise a lower bound for the number of linear measurements needed for unique reconstruction by non-convex optimization. As a result, the connected components of an image are a crucial parameter in predicting the needed number of linear measurements. Therefore, we construct an integer program model that restricts the maximum number of connected components in the reconstructed image. Further, we propose inequalities to speed-up the convergence of the branch and bound algorithm used to optimize integer programs.

In Chapter 6, we review the field of *multilevel methods* [Nas00] by analyzing unconstrained and constrained convex optimization. Then, we view the image domain as a manifold and use tools from *differential geometry* [Lee12] and *optimization on manifolds* [AMS08] to develop a first-order *multilevel* [Nas00] optimization algorithm. The developed multilevel algorithm exhibits fast convergence and enables us to recover images of higher resolution.

1.2 Publications

Parts of this thesis have already been published:

[KSP17] "A Novel Convex Relaxation for Non-binary Discrete Tomography"

[KP19] "Performance Bounds for Co-/sparse Box Constrained Signal Recovery"

In particular, Chapter 3 contains [KP19], and Chapter 4 is based on [KSP17].

CHAPTER 2

Preliminaries

In this chapter, we collect basic definitions and results that are used throughout this thesis. It is assumed that the basic terms from linear algebra and analysis are known. Proofs are generally omitted but available in the given references.

2.1 Basic Tools

2.1.1 Convex Analysis and Optimization

For a comprehensive introduction to convex analysis and convex optimization, we recommend [Roc97; RW10; Ber15].

Conventional convex optimization is based on the following definitions.

Definition 2.1 (Convex Set)**.** Let $C \subseteq \mathbb{R}^n$ be a subset of the real vector space. We call C convex if for all $x, y \in C$ and $\lambda \in [0, 1]$ it holds

$$\lambda \cdot x + (1 - \lambda) \cdot y \in C. \tag{2.1}$$

Definition 2.2 (Convex Function)**.** Let $f : C \to \mathbb{R}$ be a function and $C \subseteq \mathbb{R}^n$ a convex set. We call f convex if for all $x, y \in C$ and $\lambda \in (0, 1)$ it holds

$$f(\lambda \cdot x + (1 - \lambda) \cdot y) \leq \lambda \cdot f(x) + (1 - \lambda) \cdot f(y). \tag{2.2}$$

If the above inequality is strict ($<$), then we call f strictly convex.

Remark 2.1. Considering functions $f : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$, we call f convex if $\operatorname{dom} f$ and $f|_{\operatorname{dom} f}$ are convex.

The gradient of differentiable functions can be extended to subgradients for convex functions.

Definition 2.3 (Subdifferential)**.** Let $f : C \to \mathbb{R}$ be a convex function and $C \subseteq \mathbb{R}^n$ a convex set. Then, we call $\partial f(x)$ the subdifferential at $x \in C$ and it is defined by

$$\partial f(x) = \{g \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle g, y - x\rangle, \forall y \in \mathbb{R}^n\}. \tag{2.3}$$

Each $g \in \partial f(x)$ is called subgradient.

Remark 2.2. If f is a convex and differentiable function, then $\partial f(x) = \{\nabla f(x)\}$.

For a differentiable function f it is a necessary condition for $\overline{x} \in \mathbb{R}^n$ to be a local minimum that $\nabla f(\overline{x}) = 0$. In case f is convex and the subdifferential at $\overline{x} \in \mathbb{R}^n$ is non-empty, then $\overline{x}$ is a global minimum if and only if $0 \in \partial f(\overline{x})$. The proof can be found in [RW10, Thm. 10.1].

Within the next lemma, we provide rules calculating the subdifferential, which are frequently used in this thesis.

Lemma 2.4 (Subdifferential Calculus)**.** *Let $f, f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ for $i \in [p]$ be convex and lower semi-continuous. Then the following rules apply*

(i) if $g(x) = f(Ax)$ with $A \in \mathbb{R}^{m\times n}$ and $Ax \in \operatorname{dom} f$, then $\partial g(x) = A^\top \partial f(Ax)$;

(ii) if $f = \sum_{i\in[p]} f_i$ with $x \in \bigcap_i \operatorname{dom} f_i$, then $\partial f(x) = \sum_{i\in[p]} \partial f_i(x)$.

Proof. Using the chain rule for subdifferentials (see [RW10, Thm. 10.6]), we obtain rule (i). The second rule (ii) follows by [RW10, Cor. 10.9]. □

In the next two lemmas, we use the rules from Lemma 2.4 to calculate the subdifferential for two important functions which are used throughout this thesis.

Lemma 2.5. *Let $\overline{x} \in \mathbb{R}^n$. Then the subdifferential of the function $f(x) = \|x\|_1$ at $\overline{x}$ is defined by*

$$\partial f(\overline{x}) = \{\alpha \in \mathbb{R}^n \mid \alpha_S = \operatorname{sign}(\overline{x}_S), \|\alpha_{S^c}\|_\infty \leq 1\}, \tag{2.4}$$

where $S = \operatorname{supp}(\overline{x})$.

Proof. First, we calculate the subdifferential of the absolute value function $g(x) = |x|$

$$\forall y \in \mathbb{R} : |y| \geq |x| + \alpha \cdot y \Leftrightarrow \alpha \in \partial|x|, \tag{2.5}$$

resulting in

$$\partial|x| = \begin{cases} [-1, 1] & \text{if } x = 0, \\ \operatorname{sign}(x) & \text{otherwise.} \end{cases} \tag{2.6}$$

By Lemma 2.4, we deduce

$$\|x\|_1 = \sum_i |\langle e_i, x\rangle| \Longrightarrow \partial\|x\|_1 = \sum_i e_i \cdot \partial|x_i|. \tag{2.7}$$

Hence, the claim follows. □

Next, we consider the following constrained optimization problem

$$\min_{x \in C} f(x), \tag{2.8}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $C \subseteq \mathbb{R}^n$ a closed convex set. Using the indicator function

$$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{otherwise.} \end{cases}, \tag{2.9}$$

we transform the constrained optimization problem (2.8) to an unconstrained optimization problem by

$$\min_{x \in \mathbb{R}^n} f(x) + \delta_C(x). \tag{2.10}$$

In Chapter 3, we use formulation (2.10) and the optimality condition $0 \in \partial(f(x) + \delta_C(x))$ to obtain testable optimality conditions for a class of optimization problems. Therefore, we calculate the subdifferential of the indicator function in the following lemma.

Lemma 2.6. *Let $C \subseteq \mathbb{R}^n$ be a closed convex set and $\overline{x} \in C$. Then the subdifferential of $f(x) = \delta_C(x)$ at $\overline{x}$ is defined by*

$$\partial f(\overline{x}) = N(\overline{x}, C) = \{g \in \mathbb{R}^n \mid \langle g, y - \overline{x} \rangle \leq 0, \forall y \in C\}, \tag{2.11}$$

where $N(\overline{x}, C)$ is the normal cone of C at $\overline{x}$.

Proof. By the definition of the subdifferential, a vector $g \in \mathbb{R}^n$ is in the subdifferential $\partial f(\overline{x})$ if and only if it holds for every $y \in \mathbb{R}^n$ that

$$\delta_C(y) \geq \delta_C(\overline{x}) + \langle g, y - \overline{x} \rangle. \tag{2.12}$$

If $y \notin C$ the inequality is trivially true as $\delta_C(y) = \infty$. Thus, we only have to check if for all $y \in C$ it holds

$$0 \geq \langle g, y - \overline{x} \rangle, \tag{2.13}$$

which is the desired definition of the normal cone. □

2.1.2 Differential Geometry

Below, we provide basic terms from differential geometry, which are required in Chapter 6. For a more thorough introduction, see [Lee12]. Furthermore, we present results from [AMS08] for the optimization on smooth manifolds.

Definition 2.7 ([Lee12, p. 3])**.** A n-dimensional *manifold* is a topological space with the following properties

- $\mathcal{M}$ is **Haussdorff**: for any pair of points $p, q \in \mathcal{M}, p \neq q$, there exist disjoint open subsets $U, V \subseteq \mathcal{M}, U \cap V = \emptyset$ with $p \in U$ and $q \in V$;
- $\mathcal{M}$ is **second-countable**: there exist a countable basis for the topology of $\mathcal{M}$;
- for each $p \in \mathcal{M}$ we find
  - an open subset $U \subset \mathcal{M}$ with $p \in \mathcal{M}$,
  - an open subset $\hat{U} \subset \mathbb{R}^n$, and
  - a *homeomorphism* $\varphi : U \to \hat{U}$ called *chart*.

Remark 2.3. An *homeomorphism* f between two topological spaces is a continuous bijection such that the inverse map f^{-1} is also continuous.

In this thesis, we only consider *smooth manifolds*. Thus, we introduce the concept of an *atlas*.

Definition 2.8 ([AMS08, p. 19])**.** A family of charts (U_i, φ_i), denoted by $\mathcal{A}$, for a n-dimensional manifold $\mathcal{M}$ is called an atlas if

- $\mathcal{M} = \bigcup_i U_i$, and
- for any i, j with $U_i \cap U_j \neq \emptyset$ it holds $\varphi_i \circ \varphi_j^{-1} \in C^\infty(\mathbb{R}^n)$.

Remark 2.4. The set $C^\infty(\mathbb{R}^n)$ contains all smooth functions $f : \mathbb{R}^n \to \mathbb{R}^n$.

Definition 2.9 (Smooth Manifold)**.** Let $\mathcal{M}$ be a n-dimensional manifold. If it exists an atlas $\mathcal{A}$ for $\mathcal{M}$, then $\mathcal{M}$ is called a smooth manifold.

Remark 2.5. Henceforth, if we write manifold it means smooth manifold.

An atlas defines a differential structure on a manifold $\mathcal{M}$ so that we are able to introduce concepts like differentials and tangent vectors from classical analysis for manifolds.

We call a curve $\gamma : \mathbb{R} \to \mathcal{M}$ smooth if the composition $\varphi \circ \gamma : \mathbb{R} \to \mathbb{R}^n$ is a smooth function, where φ is a chart of $\mathcal{M}$. Using the following equivalence relation

$$\gamma_i \sim \gamma_j :\Leftrightarrow \forall (U, \varphi) : \left.\frac{\mathrm{d}}{\mathrm{d}t}(\varphi \circ \gamma_i)(t)\right|_{t=0} = \left.\frac{\mathrm{d}}{\mathrm{d}t}(\varphi \circ \gamma_j)(t)\right|_{t=0}, \tag{2.14}$$

we define the set of *tangent vectors* at a point $p \in \mathcal{M}$ as the set of equivalence classes

$$T_p\mathcal{M} = \{\gamma : \mathbb{R} \to \mathcal{M} \text{ smooth} \mid \gamma(0) = p\} / \sim . \tag{2.15}$$

Remark 2.6. The set from (2.15) forms a vector space, see [Lee12].

The *differential* of a smooth function $F : \mathcal{M} \to \mathcal{N}$ at $p \in \mathcal{M}$ between two manifolds $\mathcal{M}, \mathcal{N}$ is a map

$$\begin{aligned} \mathrm{d}_x F : & T_x\mathcal{M} \to T_{F(x)}\mathcal{N}, \\ & [\gamma] \mapsto [F \circ \gamma]. \end{aligned} \tag{2.16}$$

Proposition 2.10 ([Lee12, Prop. 3.6]). *Let $\mathcal{M}, \mathcal{N}$ and $\mathcal{P}$ be smooth manifolds, $F : \mathcal{M} \to \mathcal{N}$ and $G : \mathcal{N} \to \mathcal{P}$ be smooth maps and $p \in \mathcal{M}$. Then the differential has the following properties*

(a) $\mathrm{d}_p F : T_p\mathcal{M} \to T_{F(p)}\mathcal{N}$ *is linear;*
(b) $\mathrm{d}_p(G \circ F) = \mathrm{d}_{F(p)}G \circ \mathrm{d}_p F : T_p\mathcal{M} \to T_{G\circ F(p)}\mathcal{P}$*;*
(c) $\mathrm{d}_p(\mathrm{Id}_\mathcal{M}) = \mathrm{Id}_{T_p\mathcal{M}} : T_p\mathcal{M} \to T_p\mathcal{M}$*;*
(d) if F is a diffeomorphism, then $\mathrm{d}_p F : T_p\mathcal{M} \to T_{F(p)}\mathcal{N}$ is an isomorphism, and $(\mathrm{d}_p F^{-1}) = \mathrm{d}_{F(p)}(F^{-1})$.

Remark 2.7. A function f is a *diffeomorphism* if f and f^{-1} are continuously differentiable.

Considering $\mathcal{M} = \mathbb{R}^n$, we define a linear isomorphism between $T_p\,\mathbb{R}^n$ and $\mathbb{R}^n$ by

$$\begin{aligned} \Psi : T_p\,\mathbb{R}^n &\to \mathbb{R}^n, \\ [\gamma] &\mapsto \dot\gamma(0) := \frac{\mathrm{d}}{\mathrm{d}t}\gamma(t)\Big|_{t=0}, \\ \Psi^{-1} : \mathbb{R}^n &\to T_p\,\mathbb{R}^n, \\ v &\mapsto [\gamma(t) := p + t \cdot v]. \end{aligned} \tag{2.17}$$

Hence, any chart φ covering $p \in \mathcal{M}$ implies that $T_p\mathcal{M} \cong T_{\varphi(p)}\,\mathbb{R}^n$, see Proposition 2.10 (a) and (d). Then, we use the linear isomorphism (2.17) to identify the tangent space of a n-dimensional manifold $\mathcal{M}$ at p, i.e.,

$$T_p\mathcal{M} \cong \mathbb{R}^n\,. \tag{2.18}$$

Gradient descent is a classical first-order optimization approach to minimize a smooth function $f : \mathbb{R}^n \to \mathbb{R}$. Therefore, one computes the gradient $\nabla f(x_k)$ at some iterate $x_k \in \mathbb{R}^n$ and chooses a *descent direction* d meaning $\langle \nabla f(x_k), d\rangle < 0$. Fixing parameters $\sigma \in (0,1)$ and $\alpha^* > 0$ over all iterates, one computes $\alpha \in (0, \alpha^*]$ such that

$$f(x_k + \alpha d) \le f(x_k) + \sigma\alpha\langle \nabla f(x), d\rangle \tag{2.19}$$

and set $x_{k+1} = x_k - \alpha d$. Choosing $\alpha = \beta^m\alpha^*$ with $\beta \in (0,1)$, we seek for the smallest non-negative integer m satisfying (2.19). If the condition (2.19) is satisfied in every iteration, then the iterates provably converge to a point x^* satisfying $\nabla f(x^*) \approx 0$, see [NW06].

On manifolds, the concept of the *Armijo Line Search* is applied by introducing the Riemannian structure, which is available for every manifold [AMS08]. Thus, we equip the manifold with an inner product $\langle\cdot,\cdot\rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$. Next, we define the *Riemannian gradient* for a smooth function $f : \mathcal{M} \to \mathbb{R}$ at $x \in \mathcal{M}$, which is the unique element $\nabla_\mathcal{M} f(x) \in T_x\mathcal{M}$ such that ([AMS08, p. 46])

$$\mathrm{d}_x f[\xi] = \langle \nabla_\mathcal{M} f(x), \xi\rangle_x, \qquad \forall \xi \in T_x\mathcal{M}. \tag{2.20}$$

Similar to the classical analysis, $\xi \in T_x\mathcal{M}$ is a descent direction if and only if

$$\langle \nabla_{\mathcal{M}} f(x), \xi \rangle_x < 0. \tag{2.21}$$

On a manifold the operation $x + \alpha\xi$ for $x \in \mathcal{M}$ and $\xi \in T_x\mathcal{M}$ is not defined. Hence, we define the so called *retraction* mapping tangent vectors back to the manifold and transfers the concept of moving along a direction.

Definition 2.11 ([AMS08, Def 4.1.1])**.** A retraction on a manifold $\mathcal{M}$ is a smooth mapping $\mathfrak{R}_x : T_x\mathcal{M} \to \mathcal{M}$ with the following properties.

(i) $\mathfrak{R}_x(0_x) = x$, where 0_x denotes the zero element of $T_x\mathcal{M}$.

(ii) With the canonical identification $T_{0_x}T_x\mathcal{M} \sim T_x\mathcal{M}$, $\mathfrak{R}_x$ satisfies

$$\mathrm{d}\mathfrak{R}_x(0_x) = \mathrm{id}_{T_x\mathcal{M}}, \tag{2.22}$$

where $\mathrm{id}_{T_x\mathcal{M}}$ denotes the identity mapping on $T_x\mathcal{M}$.

To apply *Armijo Line Search* sketched in Algorithm 2.1, we need two more definitions. The first definition is needed to ensure convergence, see [AMS08, Thm. 4.3.1]. While the second is the translation of the Armijo condition to the manifold setting.

Definition 2.12 ([AMS08, Definition 4.2.1])**.** Given a cost function f on a Riemannian manifold $\mathcal{M}$, a sequence $\{\eta_k\}, \eta_k \in T_{x_k}\mathcal{M}$, is gradient-related if for any subsequence $\{x_k\}_{k\in\mathcal{K}}$ of $\{x_k\}$ that converges to a non-critical point of f, the corresponding subsequence $\{\eta_k\}_{k\in\mathcal{K}}$ is bounded and satisfies

$$\limsup_{k\to\infty, k\in\mathcal{K}} \langle \nabla_{\mathcal{M}} f(x_k), \eta_k \rangle_{x_k} < 0. \tag{2.23}$$

Definition 2.13 ([AMS08, Definition 4.2.2])**.** Given a cost function f on a Riemannian manifold $\mathcal{M}$ with retraction $\mathfrak{R}_x$, a point $x \in \mathcal{M}$, a tangent vector $\eta \in T_x\mathcal{M}$, and scalars $\overline{\alpha} > 0, \beta, \sigma \in (0, 1)$, the Armijo point is $\eta^A = t^A = \beta^m\overline{\alpha}\eta$, where m is the smallest non-negative integer such that

$$f(x) - f(\mathfrak{R}_x(\beta^m\overline{\alpha}\eta)) \geq -\sigma\langle \nabla_{\mathcal{M}} f(x), \beta^m\overline{\alpha}\eta \rangle_x. \tag{2.24}$$

The real t^A is the Armijo step size.

2.1.3 Graph Theory

Graphs are combinatorial structures which we use to index variables and model relations between variables throughout this thesis. See [KV18] for a more in-depth introduction.

Let $\mathcal{V}$ be an index set which is called *vertices* or *nodes*. Choosing a subset $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, we name each element of $\mathcal{E}$ an *edge*. If $v, w \in \mathcal{V}$ and $(v, w) \in \mathcal{E}$, then v is *adjacent* to w. Together, these sets form a *Graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. If for each $(v, w) \in \mathcal{E}$ it holds that $(w, v) \in \mathcal{E}$, we call $\mathcal{G}$ *undirected* otherwise *directed*. Henceforth, any graph is supposed

Algorithm 2.1: Armijo Line Search on Manifolds

Input: $f : \mathcal{M} \to \mathbb{R}$ is a smooth function, and $\mathfrak{R}_x : T_x\mathcal{M} \to \mathcal{M}$ is a retraction

for $k = 0, 1, \ldots$ **do**
  Pick $\eta_k \in T_{x_k}\mathcal{M}$ such that the sequence $\{\eta_i\}_{i=0,1,\ldots}$ is gradient-related
  $x_{k+1} \leftarrow R_{x_k}(t_k\eta_k)$ // t_k according to Def. 2.13
  if x_{k+1} *satisfies some optimality condition* **then**
    return x_{k+1}
  end
end

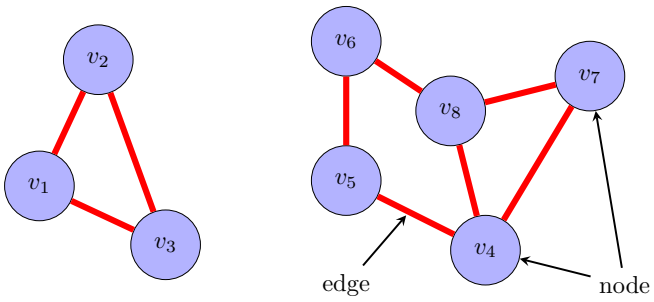


Figure 2.1: We depict a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with nodes (blue circle) and edges (red line), consisting of two connected components. A path from node v_5 to v_8 is given by $\{v_5v_6, v_6v_8\}$. If choosing the path $\{v_1v_2, v_2v_3, v_3v_1\}$, it forms a circle.

to be undirected. Keeping the notation simple, we omit the parentheses of the edges and write $vw \in \mathcal{E}$ instead of $(v, w) \in \mathcal{E}$. Further, we identify with vw and wv the same element, i.e., $vw = wv$.

A *path* between two nodes $v, w \in \mathcal{V}$ in a graph is an ordered subset

$$\{u_1u_2, u_2u_3, \ldots, u_{n-1}u_n\} \subseteq \mathcal{E} \tag{2.25}$$

such that $u_1 = v$ and $u_n = w$. If $u_1 = u_n$, we call the path a *circle*.

Choosing a subset $C \subseteq \mathcal{V}$, we call $\mathcal{G}' = (C, \mathcal{E} \cap (C \times C))$ a *subgraph* induced by C. We name C *connected* when for each pair $v, w \in C$ if it exists a path from v to w in $\mathcal{G}'$. Moreover, if it exists no set $C' \subseteq \mathcal{V}$ with $C \subseteq C'$ such that C' is connected, we call C a *connected component* of $\mathcal{G}$.

In Figure 2.1, we give a graphical illustration of a graph consisting of two connected components.

2.1.4 Branch and Bound

In chapters 4 and 5, we consider optimization problems with integrality constraints and refer to a standard approach branch and bound [Dak65; LD60] to solve them. This approach does not take a central role in this thesis. Thus, we only sketch it to

introduce the necessary terms.

Consider the following general mixed-integer optimization problem

$$\begin{aligned} \min\ & f(x,y), \\ \text{s.t. }& g(x,y) \leq 0, \\ & x \in X \cap \mathbb{Z}^p, \\ & y \in Y \cap \mathbb{R}^q, \end{aligned} \tag{MIP}$$

where f, g are continuous convex functions and $X, Y \subset \mathbb{R}^n$ are closed, convex sets. Further, we suppose that the relaxation (P) below, meaning that we drop the integrality restrictions of the variables x, is efficiently solvable and that a minimum exists.

$$\begin{aligned} \min\ & f(x,y), \\ \text{s.t. }& g(x,y) \leq 0, \\ & x \in X \cap \mathbb{R}^p, \\ & y \in Y \cap \mathbb{R}^q\,. \end{aligned} \tag{P}$$

Using branch and bound to find a solution of (MIP) is a systematic enumeration of possible integral solutions. Therefore, we iteratively solve variants of (P) and take after each iteration a *branch* or *bound* action.

The branch and bound procedure starts by solving (P) to obtain a solution x^*. If $x^* \in \mathbb{Z}^p$, we found a solution to (MIP) and the procedure stops. Otherwise there is an index $i \in [p]$ such that $x_i^* \notin \mathbb{Z}$ which results in taking the *branch* action. In the *branch* action we create two variants of the problem (P) by adding the constraint $x_i \geq \lceil x_i^* \rceil$ respectively $x_i \leq \lfloor x_i^* \rfloor$. Consequently, the current non-integral solution x^* is not feasible for either of the new variants of (P). Henceforth, we call new variants created through the *branch* action *active* problems.

Suppose that there are already active problems caused by the branch action. In each iteration of the branch and bound procedure, we remove an active problem by solving it. That results in one of the following cases,

(i) the current active problem is infeasible,
(ii) $x^* \notin \mathbb{Z}^p$, or
(iii) $x^* \in \mathbb{Z}^p$.

Regarding (i), the iteration ends and we start over as above. Facing (ii), we again apply the branch action resulting in the creation of new active problems as explained above. In the last case, we found a feasible solution pair x^*, y^* to (MIP). If this is the first feasible solution found so far, we call this solution the incumbent solution. Otherwise, we compare the new feasible solution with the current incumbent solution and set the solution with the better objective value to the new incumbent solution. Setting an incumbent solution is called the bound action. In Figure 2.2, we summarize the above steps of branch and bound in a flow-chart for a better overview.

During all iterations we keep track of the lowest objective value of all active nodes which is called the *lower bound*. The objective value of the incumbent solution is

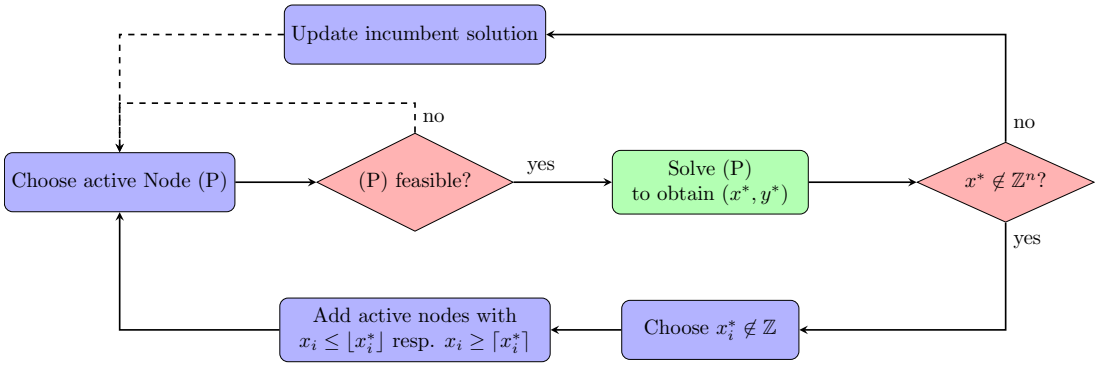


Figure 2.2: Branch and bound starts at the green node by solving the relaxation of the mixed-integer program and runs until the list of active nodes is empty or the lower bound exceeded the upper bound. Only if the list of active nodes is non-empty, we may follow the dashed arrows. Otherwise, the branch and bound procedure terminates.

called the *upper bound*. Branch and bound terminates if there is no active node left or if the lower bound exceeded the upper bound. Hence, it ends after a finite number of iterations. If the lower bound exceeded the upper bound, it implies that all other solutions obtained from the remaining active problems would yield a worse objective value. As a result, the incumbent solution after termination is the global optimal solution to (MIP).

The description above is very general, but every branch and bound implementation builds up on the idea sketched above. For a in-depth introduction, we refer to [Bal09; KV18; Sch99; Bon08; Ley01] covering the linear and non-linear case.

2.2 Compressed Sensing

This thesis focuses on the reconstruction of discrete signals, especially images, from incomplete linear measurements. Thus, we are interested in recovering a signal $\overline{x} \in \mathbb{R}^n$ undersampled by a matrix $A \in \mathbb{R}^{m \times n}$ with $m \ll n$ from linear measurements

$$Ax = A\overline{x} =: b. \tag{2.26}$$

Assuming a full rank matrix A, the linear system above has infinitely many solutions. Hence, we cannot reconstruct $\overline{x}$ from the linear system (2.26) without additional information. For this reason, we incorporate in the reconstruction model that the signal meets a sparsity condition. There is a wide-ranging theory reconstructing such signals underlying a low-complexity model. This theory is known as *Compressed Sensing* (CS) and provides mathematically substantiated conditions for signal recovery from undersampled data. An in-depth introduction can be found in [FR13]. We only give a short introduction as chapters 3 and 5 cover, and extend further topics of CS.

Definition 2.14 (Sparsity)**.** Let $x \in \mathbb{R}^n$ be a vector. Then we call x a s-sparse vector with $s \in \mathbb{N}$ if

$$\|x\|_0 := |\operatorname{supp}(x)| \leq s, \tag{2.27}$$

where $\operatorname{supp}(x) = \{i \in [n] \ : \ x_i \neq 0\}$.

Knowing that the desired signal $\overline{x} \in \mathbb{R}^n$ is s-sparse so that s is small compared to n, we only seek for solutions of the given linear system which are sparse. As a result, CS examines the solutions of the following combinatorial optimization problem

$$\min \|x\|_0 \text{ s.t. } Ax = b := A\overline{x}. \tag{P_0}$$

Ensuring that each s-sparse solution is unique, a necessary and sufficient condition for A is that

$$\operatorname{spark}(A) := \min\{\|v\|_0 \mid v \in \mathcal{N}(A) \setminus \{0\}\} > 2s. \tag{2.28}$$

As a consequence, if the matrix A satisfies the condition (2.28) and the desired signal $\overline{x}$ is s-sparse, then $\overline{x}$ can be uniquely recovered by (P_0). Chapter 5 introduces a more general theory called *union of subspaces* [LD08; BD09] implying the result (2.28).

Both the optimization of (P_0) and the calculation of (2.28) are known to be NP-hard, see [Nat95; TP14]. Therefore, CS considers the convex relaxation

$$\min \|x\|_1 \text{ s.t. } Ax = b, \tag{BP}$$

which is called *basis pursuit*. Similar to the condition that involves the spark, there is also a necessary and sufficient condition for (BP) to guarantee reconstruction of any s-sparse signal.

Definition 2.15 (Null Space Property)**.** Let $A \in \mathbb{R}^{m \times n}$. Then A is said to have the *Null Space Property* (NSP) relative to a set $S \subset [n]$ if

$$\|v_S\|_1 < \|v_{S^c}\|_1, \qquad \forall v \in \mathcal{N}(A) \setminus \{0\}. \tag{2.29}$$

The matrix A is said to have the NSP of order s if

$$\|v_S\|_1 < \|v_{S^c}\|_1, \qquad \forall v \in \mathcal{N}(A) \setminus \{0\}, \ \forall S \subset [n] : |S| \leq s. \tag{2.30}$$

Chapter 3 focuses on the reconstruction of signals through a generalized form of (BP), where we replace $\|x\|_1$ by $\|\Omega x\|_1$ with $\Omega \in \mathbb{R}^{p \times n}$. This generalized problem requires also a general form of the *null space property*.

In particular, we are interested if Ω is the two-dimensional discrete gradient denoted by ∇. For an one-dimensional signal $x \in \mathbb{R}^d$ the discrete gradient Dx is defined by

$$Dx = \begin{pmatrix} x_2 - x_1 \\ \vdots \\ x_n - x_{d-1} \end{pmatrix} \in \mathbb{R}^{d-1}. \tag{2.31}$$

Using the one-dimensional discrete gradient, we define the two-dimensional version by

$$\nabla = \begin{pmatrix} I \otimes D \\ D \otimes I \end{pmatrix}, \tag{2.32}$$

where $\otimes$ is the Kronecker product. For the two-dimensional discrete gradient, we assume that a two-dimensional signal from $\mathbb{R}^{d\times d}$ is stacked column-wise into a vector $x \in \mathbb{R}^n$, where $n = d^2$. Recovering gradient sparse two-dimensional signals, we use

$$\min \|\nabla x\|_1 \text{ s.t. } Ax = b. \tag{2.33}$$

The main interest of this thesis is to investigate uniform recovery guarantees for (2.33), when A is a tomographic matrix (see subsequent Section 2.3) However, the results from CS only apply to sparse and not gradient sparse signals. One result which extends a uniform recovery guarantee from CS to gradient sparse images is from [NW13b]. They use the Haar wavelet transform and the *restricted isometry property* (RIP) to guarantee unique recovery for gradient sparse images.

Definition 2.16. Let $A \in \mathbb{R}^{m\times n}$ be a matrix and $\delta \in (0, 1)$. The matrix A has the RIP of order s for $s \in [n]$, if for all s-sparse vectors $x \in \mathbb{R}^n$ it holds

$$(1 - \delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta)\|x\|_2^2. \tag{2.34}$$

The smallest δ denoted by δ_s satisfying (2.34) is called the RIP constant.

Since the result of [NW13b] applies also to continuous images, we provide the discretized version from [Den14a]. Then, the results reads that if the matrix AH^{-1} has a RIP constant $\delta_{5s} < \frac{1}{3}$, we uniquely recover every s gradient sparse signal. With H we denote the discrete Haar wavelet transform.

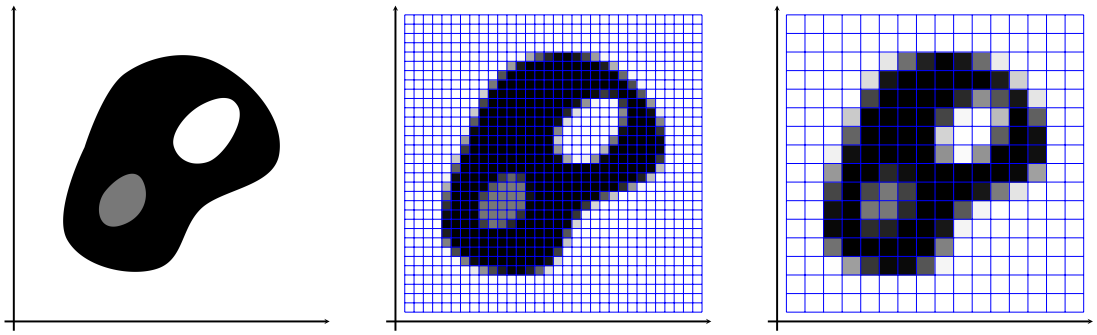
Considering tomographic matrix, the authors of [Den14a] showed that using the Haar wavelet transform, it only guarantees the reconstruction of images with low gradient sparsity. In Chapter 3, we see that tomographic matrices perform much better meaning that even with few projections we are able to recover images with larger sparsity levels of the gradient.

2.3 Discrete Tomography

In this section, we define the projection geometries of *discrete tomography* which are used throughout this thesis. For the physics of computer tomography, we refer to [Buz08; Her09]. An overview of *discrete tomography* can be found in [HK07].

The projection geometries used in this thesis base on the X-ray transform model [Nat01], closely related to the Radon transform, which maps a function representing an image $u : \mathbb{R}^2 \to \mathbb{R}$ to its line integral

$$f(s, p) = \int u(s + tp)\mathrm{d}t, \tag{2.35}$$



(a) continous domain (b) 32×32 pixel grid (c) 16×16 pixel grid

Figure 2.3: In (a), we depicted an image with a continuous domain. Using a finite grid of pixel basis functions, we receive in (b) and (c) different discretizations. The gray value of each square represents the coefficient of a pixel basis function at the spatial position.

where $s, p \in \mathbb{R}^2$ are linear independent vectors. Such a transform is a continuous-to-continuous operation, since we map a function to a new function. The filtered backprojection, mentioned in the introduction of this thesis, is an inversion formula of (2.35) to retrieve the function u if only f is known.

In *discrete tomography*, we assume that the image function is discretized by a finite expansion set, e.g., pixel, and $u \in \mathbb{R}^n$ are the coefficients. Figure 2.3 illustrates the approximation of a continuous image domain by a finite pixel grid. Let $\mathcal{B}_j$ be the j-th basis function. Then, for a single ray $s_i + t \cdot p_i$ the transform (2.35) rewrites to

$$b_i = \sum_{i \in [n]} u_i \underbrace{\int \mathcal{B}_j(s_i + t \cdot p_i) \mathrm{d}t}_{=:A_{ij}} . \tag{2.36}$$

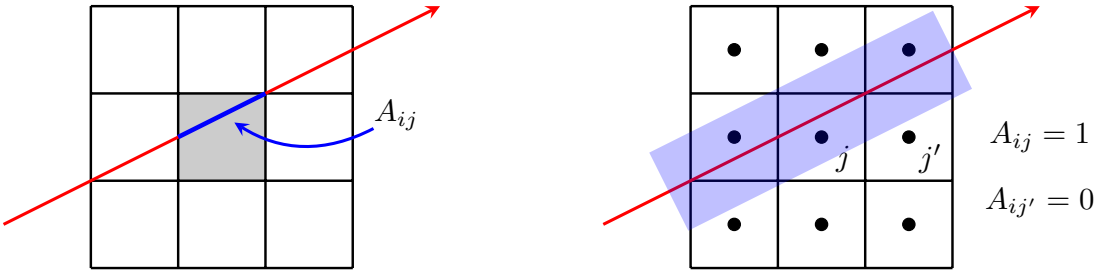
Collecting the coefficients A_{ij} from (2.36) into a matrix, we obtain a discrete-to-discrete transformation in contrast to (2.35) by

$$b = Au, \tag{2.37}$$

where $b \in \mathbb{R}^m$ are the result of choosing a finite set of m rays and $A \in \mathbb{R}^{m \times n}$. Hence, the entries of A provides an incidence geometry defining how the basis function j contributes to the i-th X-ray.

Throughout this thesis, we consider two basis functions to construct A, which are presented in the remaining of this section. Henceforth, we consider the image to be discretized by pixels on a $d \times d$ grid. We identify each pixel by an index $j \in [d^2]$.

Let $\mathcal{B}_j$ be the basis function corresponding to the pixel j, meaning that it yields one



(a) real weights (b) binary weights

Figure 2.4: The left figure shows real weight from (2.38) and on the right we see the weights from (2.39). We marked the region of points with distance less than $\frac{1}{2}$ to the ray in blue.

in the square defined by the pixel and zero otherwise. Then,

$$A_{ij} = \int \mathcal{B}_j(s_i + t \cdot p_i)\mathrm{d}t \tag{2.38}$$

is the intersection length of ray i with pixel j. In this case, we call the entries of the tomographic matrix A *real weights*. Figure 2.4 (a) illustrates that intersection of pixel j and ray i to obtain (2.38).

The authors in [RLH14] presented *binary weights*, i.e., $A_{ij} \in \{0, 1\}$, to model A. Therefore, we define that the index $j \in [d]^2$ represents the center of pixel j. Then, we replace the basis function in (2.36) by the Heaviside[1] step function to decide whether the distance of the center of pixel j to ray i is less than $\frac{1}{2}$,

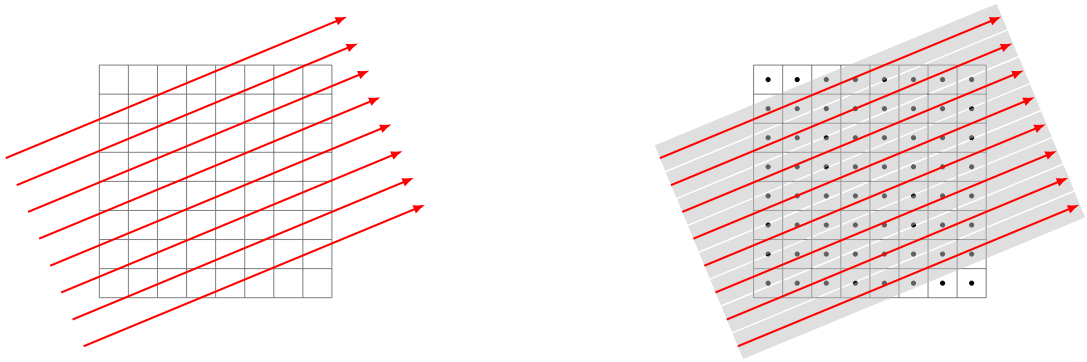
$$A_{ij} = H\left(\frac{1}{2} - \frac{|\langle j - s_i, s_i\rangle|}{\|s_i\|_2}\right). \tag{2.39}$$

In Figure 2.4 (b), we marked the region of points with a distance smaller than $\frac{1}{2}$ to ray i in blue to show which pixel assigned with weight zero or one.

So far, we explained the construction of the weights along a ray. We call a set of parallel rays a *projection* which are uniformly distributed around the center of the image domain. Figure 2.5 depicts projections for real or binary choice of weights.

Throughout this thesis, we keep the number of rays in a projection constant and obtain new projections by rotating the rays simultaneously around the image center which was proposed in [RLH14]. This is illustrated in Figure 2.6. Keeping the number of rays constant over all projections, we make sure that each projection collects approximately the same amount of information.

[1] $H(x)$ is one if $x \geq 0$ and zero otherwise.



(a) parallel real (b) parallel binary

Figure 2.5: Both figures show a set of parallel rays which are uniformly distributed around the image center. In the left figure, we choose real weights and binary weights in the right image.

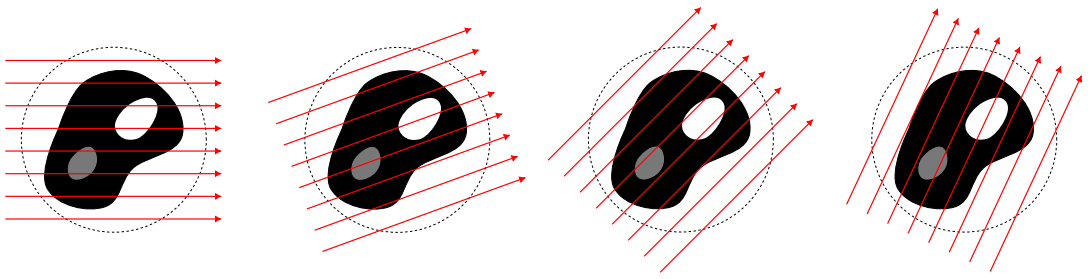


Figure 2.6: In this figure, we show four different projections which are counter-clockwise rotated from left to right. With the dashed circle, we mark the domain of the image which nonzero values. The projections are rotated around the center of the dashed circle.

CHAPTER 3

Box-Constrained Sparse and Cosparse Recovery

This chapter is motivated by the image reconstruction problem from few linear measurements, more specifically by the tomographic recovery [HK99] of industrial objects. Dealing with the severe under-sampling of an image, we want to exploit prior knowledge like the gradient sparsity and the small finite range of signal values. In particular, we are concerned with the recovery of box-constrained gradient sparse signals.

Recovering a structured signal/image from few linear measurements is a central point in both compressed sensing (CS) [FR13] and discrete tomography [HK99]. In CS, the signal structure is described through a low complexity model. For instance, if the signal $x \in \mathbb{R}^n$ is sparse, i.e., $\|x\|_0 =: s \ll n$, the associated ℓ_0-regularization is typically relaxed to the sparsity promoting convex ℓ_1-regularization. The theory of CS implies that if the number of measurements $m \geq C \cdot s \cdot \log(n/s)$ and the entries of the measurement matrix A follows a normal distribution, i.e., $a_{ij} \sim \mathcal{N}(0,1)$, the solution of ℓ_0-minimization is equal to the solution of ℓ_1-minimization. Hence, a combinatorial problem is solved by a convex optimization problem.

In the following, we consider the noiseless setting

$$A\overline{x} = b, \tag{3.1}$$

where the unknown (structured) signal $\overline{x} \in \mathbb{R}^n$ is under-sampled either by a Gaussian matrix $A \in \mathbb{R}^{m \times n}$, i.e., A has independent standard normal entries, or by a tomographic projection matrix A.

For the reconstruction of the sparse or gradient sparse box-constrained signal of interest, we consider some structure enforcing regularizer $f : \mathbb{R}^n \to \mathbb{R}$, that is a proper convex function, and solve

$$\min f(x) \text{ s.t. } Ax = A\overline{x}, \tag{3.2}$$

where we consider f as one of the following functions

$$f_1(x) = \|x\|_1, \tag{3.2a}$$
$$f_2(x) = \|x\|_1 + \delta_{\mathbb{R}^n_{\geq 0}}(x), \tag{3.2b}$$
$$f_3(x) = \|x\|_1 + \delta_{[0,1]^n}(x), \tag{3.2c}$$
$$f_4(x) = \|\Omega x\|_1, \tag{3.2d}$$
$$f_5(x) = \|\Omega x\|_1 + \delta_{\mathbb{R}^n_{\geq 0}}(x), \tag{3.2e}$$
$$f_6(x) = \|\Omega x\|_1 + \delta_{[0,1]^n}(x). \tag{3.2f}$$

In the first part of this chapter, $\Omega \in \mathbb{R}^{p\times n}$ is an arbitrary matrix, while in the latter part, Ω corresponds to the discrete gradient operator, in view of the (anisotropic) TV-regularizer we are interested in.

3.1 Individual Recovery

In this section, we derive testable uniqueness conditions for the considered problems (3.2a) - (3.2f). We first provide recovery conditions called dual certificates. More precisely, considering that given a specific vector x, and we have to decide whether it is the unique solution of (3.2). We provide necessary and sufficient conditions that certify the existence and uniqueness of a solution (3.2) for the case of a polyhedral function f. Our analysis closely follows [Gil16]. These conditions are formulated in terms of a solution to the dual problem of (3.2). For the cases (3.2a) - (3.2f), we see that it is possible to test uniqueness by merely solving a linear program. In the following, it is useful to recast (3.2) as an unconstrained optimization problem

$$\min_{x\in\mathbb{R}^n} g(x) := f(x) + \delta_{\mathcal{X}}(x), \tag{3.3}$$

where $\mathcal{X}$ denotes the feasible set of (3.2) with

$$\mathcal{X} := \{x \in \mathbb{R}^n \mid Ax = A\overline{x}\}. \tag{3.4}$$

Hence, the function g considered in this section is convex polyhedral, meaning that its epigraph is a convex polyhedron.

3.1.1 Dual Certificates

Hereafter, we use the result from [Gil16] giving us a necessary and sufficient condition for x to be the unique minimum of a polyhedral function.

Lemma 3.1 ([Gil16, Lem. 2.2]). *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be any proper, convex and lower semi-continuous function that is also polyhedral[1]. Then it holds*

$$\{\overline{x}\} = \arg\min_x g(x) \Leftrightarrow 0 \in \operatorname{int}(\partial g(\overline{x})). \tag{3.5}$$

The main result, Theorem 3.3, in this section, concerns the uniqueness of the optimization problem

$$\min_{x\in\mathbb{R}^n} \|\Omega x\|_1 + \delta_{\mathcal{X}}(x) + \delta_{[l,u]^n}(x). \tag{3.6}$$

It is a generalization of (3.2a) - (3.2f) since one sets $\Omega = I$, $l = -\infty$, or $u = \infty$ to achieve any of the before mentioned optimization problems.

Definition 3.2. Let $x, l, u \in \mathbb{R}^n$ with $l < u$, and $\Psi : \mathbb{R}^n \to \{-1, 0, 1\}^{n\times n}$. The function Ψ maps a vector to a diagonal matrix, which is defined as

$$(\Psi(x, l, u))_{ii} := \begin{cases} -1 & \text{if } l_i = x_i \\ 1 & \text{if } u_i = x_i \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

If it is apparent from context, we write Ψ instead of $\Psi(x, l, u)$.

Theorem 3.3 (Dual Certificate). *A solution $\overline{x}$ to (3.6) is unique if and only if the following conditions hold*

(i) $\mathcal{N}(A) \cap \mathcal{N}(\Omega_{\Lambda,\bullet}) \cap \mathcal{N}(\Psi) = \{0\}$

(ii) $\exists \alpha \in \mathbb{R}^p,\ \exists \mu \in \mathbb{R}^n$:,
$\Omega^\top \alpha + \Psi\mu \in \mathcal{R}(A^\top),\ \alpha_{\Lambda^c} = \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}),\ \|\alpha_\Lambda\|_\infty < 1,\ \mu > 0,$

where $\Lambda := \operatorname{cosupp}(\Omega\overline{x})$ *and* Ψ *from Definition 3.2.*

Proof. By Lemma 3.1, $\overline{x}$ is the unique solution to (3.6) if and only if

$$0 \in \operatorname{int}(\partial(\|\Omega\overline{x}\|_1 + \delta_{\mathcal{X}}(\overline{x}) + \delta_{[l,u]}(\overline{x}))). \tag{3.8}$$

We rewrite (3.8) into:

(a) $\mathbb{R}^n = \operatorname{aff}(\partial(\|\Omega\overline{x}\|_1 + \delta_{\mathcal{X}}(\overline{x}) + \delta_{[l,u]}(\overline{x})))$,
(b) $0 \in \operatorname{relint}(\partial(\|\Omega\overline{x}\|_1 + \delta_{\mathcal{X}}(\overline{x}) + \delta_{[l,u]}(\overline{x})))$.

Condition (a) ensures that the interior of the subdifferential is non-empty, and is converted to (see Lemma 2.4)

$$\mathbb{R}^n = \operatorname{aff}(\partial\|\Omega\overline{x}\|_1) + \operatorname{aff}(\partial\delta_{\mathcal{X}}(\overline{x})) + \operatorname{aff}(\partial\delta_{[l,u]}(\overline{x}))). \tag{3.9}$$

Applying the subdifferential calculus from Section 2.1.1, we calculate the individual

[1] A function is polyhedral if it is piecewise affine linear.

subdifferentials from (3.9) and the affine hull

$$\begin{aligned}
\operatorname{aff}(\partial\|\Omega\overline{x}\|_1) &= \operatorname{aff}(\{\Omega^\top\alpha \mid \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}) = \alpha_{\Lambda^c}, \|\alpha_\Lambda\|_\infty \leq 1\}) \\
&= \underbrace{\Omega^\top_{\Lambda^c,\bullet}\operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x})}_{=:y_0} + \operatorname{aff}(\{\Omega^\top_{\Lambda,\bullet}\alpha \mid \|\alpha\|_\infty \leq 1\}) \\
&= y_0 + \mathcal{R}(\Omega^\top_{\Lambda,\bullet}), \\
\operatorname{aff}(\delta_{\mathcal{X}}(\overline{x})) &= \operatorname{aff}(N(\overline{x},\mathcal{X})) = \mathcal{N}(A)^\perp, \\
\operatorname{aff}(\delta_{[l,u]}) &= \operatorname{aff}(N(\overline{x},[l,u]) = \operatorname{aff}(\mathbb{R}_-^{|S_l|} \times \{0\}^{|S_{lu}|} \times \mathbb{R}_+^{S_u}) \\
&= \mathcal{R}(\Psi),
\end{aligned}$$

where $S_l = \{i \mid \overline{x}_i = l_i\}, S_u = \{i \mid \overline{x}_i = u_i\}$ and $S_{lu} = \{i \mid l_i < \overline{x}_i < u_i\}$. Consequently, we transform (a) to

$$\mathbb{R}^n = y_0 + \mathcal{R}(\Omega^\top_{\Lambda,\bullet}) + \mathcal{N}(A)^\perp + \mathcal{R}(\Psi). \tag{3.10}$$

Consider that $\mathbb{R}^n + v = \mathbb{R}^n$ for all $v \in \mathbb{R}^n$, we omit y_0 in the equation above. Then, taking the orthogonal complement without y_0, we obtain

$$\begin{aligned}
\{0\} = (\mathbb{R}^n)^\perp &= (\mathcal{R}(\Omega^\top_{\Lambda,\bullet}) + \mathcal{N}(A)^\perp + \mathcal{R}(\Psi))^\perp && (3.11)\\
&= (\mathcal{R}(\Omega^\top_{\Lambda,\bullet}))^\perp \cap (\mathcal{N}(A)^\perp)^\perp \cap \mathcal{R}(\Psi)^\perp && (3.12)\\
&= \mathcal{N}(\Omega_{\Lambda,\bullet}) \cap \mathcal{N}(A) \cap \mathcal{R}(\Psi), && (3.13)
\end{aligned}$$

which yields (i). Next we reformulate condition (b). For this purpose, we use the result [Roc97, Cor. 6.6.2] stating that for two convex sets C_1, C_2 it holds

$$\operatorname{relint}(C_1 + C_2) = \operatorname{relint}(C_1) + \operatorname{relint}(C_2). \tag{3.14}$$

Employing (3.14) and the definition of the subdifferential (see Def. 2.3), we obtain

$$\begin{aligned}
0 \in\ & \operatorname{relint}(\partial(\|\Omega\overline{x}\|_1 + \delta_{\mathcal{X}}(\overline{x}) + \delta_{[l,u]}(\overline{x}))) \\
=\ & \operatorname{relint}(\partial(\|\Omega\overline{x}\|_1) + \operatorname{relint}(\delta_{\mathcal{X}}(\overline{x})) + \operatorname{relint}(\delta_{[l,u]}(\overline{x}))) \\
=\ & \operatorname{relint}(y_0 + \{\Omega^\top_{\Lambda,\bullet}\alpha \mid \|\alpha\|_\infty \leq 1\}) + \operatorname{relint}(\mathcal{N}(A)^\perp) + \operatorname{relint}(\{\Psi y \mid y \geq 0\}) \\
=\ & \operatorname{relint}(y_0 + \{\Omega^\top_{\Lambda,\bullet}\alpha \mid \|\alpha\|_\infty \leq 1\}) + \operatorname{relint}(\mathcal{R}(A^\top)) + \operatorname{relint}(\{\Psi y \mid y \geq 0\}) \\
=\ & y_0 + \{\Omega^\top_{\Lambda,\bullet}\alpha \mid \|\alpha\|_\infty < 1\} + \mathcal{R}(A^\top) + \{\Psi y \mid y > 0\},
\end{aligned}$$

implying (ii), which concludes the proof. □

Below, we demonstrate applications of the Theorem 3.3 to (3.2a) - (3.2f) as corollaries. In Section 3.1.3, we reveal that the conditions from Theorem 3.3 are efficiently checked by linear programming.

Corollary 3.4. *A solution* $\overline{x}$ *of* (3.2a) *is the unique if and only if*

$$\exists y \in \mathbb{R}^m : A_{\bullet,S}^\top y = \operatorname{sign}(\overline{x}_S) \wedge \|A_{\bullet,S^c}^\top y\|_\infty < 1 \text{ and } A_{\bullet,S} \text{ is injective,}$$

where $S := \operatorname{supp}(\overline{x})$.

Proof. Setting $\Omega = I$, $l = -\infty$ and $u = \infty$, we achieve (3.2a) from (3.6) such that we apply Theorem 3.3. Consequently, we obtain $\Psi = 0$ such that the conditions from Theorem 3.3 are reduced to

(i) $\mathcal{N}(A) \cap \mathcal{N}(I_{S^c,\bullet}) = \{0\}$,
(ii) $\exists \alpha \in \mathbb{R}^n : \alpha \in \mathcal{R}(A^\top), \alpha_S = \operatorname{sign}(\overline{x}_S), \|\alpha_{S^c}\|_\infty < 1$,

taking into account that $\Lambda = S^c$ and $\mathcal{N}(\Psi) = \mathbb{R}^n$. Condition (i) is further simplified to

$$\begin{aligned}\{0\} = \mathcal{N}(A) \cap \mathcal{N}(I_{S^c,\bullet}) &= \mathcal{N}(A) \cap \mathcal{N}(I_{\bullet,S^c}) \\ &= \mathcal{N}(A) \cap \{x \colon x_{S^c} = 0\} \\ &= \mathcal{N}(A_{\bullet,S}),\end{aligned}$$

which yields the condition that $A_{\bullet,S}$ has to be injective. We conclude the proof by transforming the condition (ii) to

$$\exists \alpha \in \mathbb{R}^n, y \in \mathbb{R}^m : \alpha = A^\top y \wedge \alpha_S = \operatorname{sign}(\overline{x}_S) \wedge \|\alpha_{S^c}\|_\infty < 1 \tag{3.15}$$

$$\Leftrightarrow \exists y \in \mathbb{R}^m : A_{\bullet,S}^\top y = \operatorname{sign}(\overline{x}_S) \wedge \|A_{\bullet,S^c}^\top y\|_\infty < 1. \tag{3.16}$$

□

Corollary 3.5. *A solution* $\overline{x}$ *of* (3.2b) *is the unique if and only if*

$$\exists y \in \mathbb{R}^m : A_{\bullet,S}^\top y = \mathbb{1}_S \wedge A_{\bullet,S^c}^\top y < \mathbb{1}_{S^c} \text{ and } A_{\bullet,S} \text{ is injective,}$$

with $S := \operatorname{supp}(\overline{x})$.

Proof. In order to apply Theorem 3.3, we proceed similarly to the proof of Corollary 3.4 by setting $\Omega = I$, $l = 0$, and $u = \infty$, resulting in the conditions

(i) $\mathcal{N}(A) \cap \mathcal{N}(I_{S^c,\bullet}) \cap \mathcal{N}(-I_{S^c,\bullet}) = \{0\}$,
(ii) $\exists \alpha \in \mathbb{R}^n, \exists \mu \in \mathbb{R}^n :$
$\alpha - I_{S^c,\bullet}\mu \in \mathcal{R}(A^\top), \alpha_S = \operatorname{sign}(\overline{x}_S), \|\alpha_{S^c}\|_\infty < 1, \mu > 0.$

where $\Lambda = S^c$ and $\Psi = -I_{S^c,\bullet}$. Likewise to Corollary 3.4, condition (i) yields that $A_{S,\bullet}$ has to be injective.

Condition (ii) rewrites to

$$A_{\bullet,S}^\top y = \mathbb{1}_S \text{ and } A_{\bullet,S^c}^\top y + \mu_{S^c} = \alpha_{S^c} \tag{3.17}$$

with $\|\alpha_{S^c}\|_\infty < 1$ and $\mu_{S^c,\bullet} > 0$. Hence, we rewrite the second part of (3.17) as

$$A^\top_{\bullet,S^c} y < \mathbb{1}_{S^c}, \tag{3.18}$$

which concludes the proof. □

Corollary 3.6. *A binary solution $\overline{x} \in \{0,1\}^n$ of (3.2c) is the unique if and only if*

$$\exists y \in \mathbb{R}^m \;:\; A^\top_{\bullet,S} y > \mathbb{1} \wedge A^\top_{\bullet,S^c} y < \mathbb{1}, \tag{3.19}$$

with $S := \operatorname{supp}(\overline{x})$.

Proof. Like in the previous corollaries, we set $\Omega = I$, $l = 0$ and $u = 1$. As a result, the matrix Ψ has only nonzero entries on its diagonal, which leads to $\mathcal{N}(\Psi) = \{0\}$. Consequently, condition (i) of Theorem 3.3 is fulfilled. Similar to Corollary 3.5, the condition (ii) of Theorem 3.3 rewrites to

$$A^\top_{\bullet,S} y - \mu_S = \mathbb{1}_S \tag{3.20}$$

$$\text{and } A^\top_{\bullet,S^c} y + \mu_{S^c} = \alpha_{S^c} \tag{3.21}$$

with $\|\alpha_{S^c}\|_\infty < 1$ and $\mu > 0$. By transforming the above equations to

$$A^\top_{\bullet,S} y > \mathbb{1}_S \tag{3.22}$$

$$\text{and } A^\top_{\bullet,S^c} y < \mathbb{1}_S, \tag{3.23}$$

we conclude the proof. □

Corollary 3.7. *A solution $\overline{x} \in \mathbb{R}^n$ of (3.2d) is the unique if and only if*

(i) $\mathcal{N}(A) \cap \mathcal{N}(\Omega_{\Lambda,\bullet}) = \{0\}$,

(ii) $\exists \alpha : \Omega^\top \alpha \in \mathcal{R}(A^\top), \alpha_{\Lambda^c} = \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}), \|\alpha_\Lambda\|_\infty < 1$,

where $\Lambda = \operatorname{cosupp}(\Omega\overline{x})$.

Proof. As in the proof of Corollary 3.4, we conclude that $\Psi = 0$ is the zero matrix. Hence, the two conditions of Theorem 3.3 simplify to the conditions above, that do not involve the variable μ. □

Corollary 3.8. *A solution $\overline{x} \in \mathbb{R}^n_+$ of (3.2e) is the unique if and only if*

(i) $\mathcal{N}(A_S) \cap \mathcal{N}(\Omega_{\Lambda,S}) = \{0\}$,

(ii) $\exists \alpha, \exists \mu : \Omega^\top \alpha - I_{\bullet,S}\mu \in \mathcal{R}(A^\top), \alpha = \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}), \|\alpha_\Lambda\|_\infty < 1, \mu > 0$

where $\Lambda = \operatorname{cosupp}(\Omega\overline{x})$ and $S = \operatorname{supp}(\overline{x})$.

Proof. Immediate from Theorem 3.3 and Corollary 3.4 considering that $\Psi = -I_{\bullet,S}$. □

Corollary 3.9. *A binary solution $\overline{x} \in \{0,1\}^n$ of (3.2f) is the unique if and only if*

$$\exists \alpha, \exists \mu > 0 : \Omega^\top \alpha + \Psi\mu \in \mathcal{R}(A^\top), \alpha = \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}), \|\alpha_\Lambda\|_\infty < 1,$$

where $\Lambda = \operatorname{cosupp}(\Omega\overline{x})$.

Proof. Like in Corollary 3.6, the condition (i) of Theorem 3.3 is always valid. Consequently, we just have to consider condition (ii) of Theorem 3.3. □

3.1.2 Null Space Property

In the preliminaries, we have already introduced the null space property for basis pursuit. We show in the following theorem that an extended version of the null space property is equivalent to Theorem 3.3.

Theorem 3.10. *The optimality conditions in Theorem 3.3 are equivalent to the box-constrained null space property*

$$\langle \Omega_{\Lambda^c,\bullet} v, \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x})\rangle < \|\Omega_{\Lambda,\bullet} v\|_1, \qquad \forall v \in \mathcal{N}(A)\setminus\{0\}, \Psi v \geq 0. \tag{3.24}$$

Proof. We start by assuming that (3.24) holds and show that $\overline{x}$ is the unique minimizer to (3.6). Recall that uniqueness is equivalent to the conditions of Theorem 3.3. In the following, we will choose an arbitrary $\overline{x} \neq y \in [l,u]$ with $Ay = A\overline{x}$. It is straightforward to see that $\Psi(\overline{x}-y) \geq 0$, where Ψ is defined as in Theorem 3.3.

$$\begin{aligned}
\|\Omega\overline{x}\|_1 &= \langle \Omega_{\Lambda^c,\bullet}\overline{x}, \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x})\rangle \\
&\overset{①}{=} \langle \Omega_{\Lambda^c,\bullet}\overline{x} - \Omega_{\Lambda^c,\bullet} y + \Omega_{\Lambda^c,\bullet} y, \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x})\rangle \\
&= \langle \Omega_{\Lambda^c,\bullet}(\overline{x}-y), \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x})\rangle + \langle \Omega_{\Lambda^c,\bullet} y, \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x})\rangle \\
&\overset{②}{<} \|\Omega_{\Lambda,\bullet}(\overline{x}-y)\|_1 + \|\Omega_{\Lambda^c,\bullet} y\|_1 \\
&\overset{③}{=} \|\Omega_{\Lambda,\bullet} y\|_1 + \|\Omega_{\Lambda^c,\bullet} y\|_1 \\
&= \|\Omega y\|_1.
\end{aligned}$$

We obtain ① through a zero expansion. Using $\langle \Omega_{\Lambda^c,\bullet} y, \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x})\rangle \leq \|\Omega_{\Lambda^c,\bullet} y\|_1$ and (3.24), we achieve ②. Finally, we use $\Omega_{\Lambda,\bullet}\overline{x} = 0$ to gain ③. Consequently, $\|\Omega\overline{x}\|_1 < \|\Omega y\|_1$ holds for all $\overline{x} \neq y \in [l,u]$, which implies that $\overline{x}$ is a unique minimizer.

Conversely, we suppose that $\overline{x}$ is a unique minimizer of (3.6). According to Theorem 3.3, we find α with $\|\alpha\|_\infty < 1$ and $\mu > 0$ such that

$$\begin{aligned}
&\langle v, \underbrace{\Omega_{\Lambda^c,\bullet}^\top \alpha + \Omega_{\Lambda^c,\bullet}^\top \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}) + \Psi\mu}_{\in \mathcal{N}(A)^\perp}\rangle = 0 && v \in \mathcal{N}(A) \\
\Leftrightarrow\ &\langle v, \Omega_{\Lambda^c,\bullet}^\top \alpha + \Omega_{\Lambda^c,\bullet}^\top \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x})\rangle = -\langle v, \Psi\mu\rangle && v \in \mathcal{N}(A).
\end{aligned}$$

Further, we only consider $v \in \mathcal{N}(A)$ with $\Psi v \geq 0$. This leads to $\langle \Psi v, \mu\rangle \geq 0$, since $\mu > 0$. Consequently, we obtain

$$\begin{aligned}
&\langle \Omega_{\Lambda^c,\bullet} v, \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x})\rangle \leq -\langle \Omega_{\Lambda,\bullet} v, \alpha\rangle, && \forall v \in \mathcal{N}(A)\setminus\{0\} \text{ with } \Psi v \geq 0 \\
\Leftrightarrow\ &\langle \Omega_{\Lambda^c,\bullet} v, \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x})\rangle \leq \|\Omega_{\Lambda,\bullet} v\|_1, && \forall v \in \mathcal{N}(A)\setminus\{0\} \text{ with } \Psi v \geq 0.
\end{aligned}$$

The above inequality becomes strict if $\Omega_{\Lambda,\bullet} v \neq 0$ or $\Psi v \neq 0$. Condition (i) from

Theorem 3.3 ensures that $\Omega_{\Lambda,\bullet}v$ and Ψv can not simultaneously be equal to the zero vector. Hence (3.24) holds. □

Remark 3.1. In [Nam13, Thm. 7] the authors showed a similar result to Theorem 3.10, but without considering box constraints.

3.1.3 Calculating the Dual Certificate

Next, we discuss how the conditions of Theorem 3.3 are verified using linear programming. To this end, let $\overline{x}$ be a solution of (3.6). We need to check the conditions

(i) $\mathcal{N}(A) \cap \mathcal{N}(\Omega_{\Lambda,\bullet}) \cap \mathcal{N}(\Psi) = \{0\}$
(ii) $\exists \alpha \in \mathbb{R}^p$, $\exists \mu \in \mathbb{R}^n$:
$\Omega^\top \alpha + \Psi\mu \in \mathcal{R}(A^\top),\ \alpha_{\Lambda^c} = \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}),\ \|\alpha_\Lambda\|_\infty < 1,\ \mu > 0,$

from Theorem 3.3. Condition (i) is equivalent to test if

$$\operatorname{rank}\begin{pmatrix} A \\ \Omega_{\Lambda,\bullet} \\ \Psi \end{pmatrix} = n. \tag{3.25}$$

The second condition (ii) could be verified in practice by minimizing $\|\alpha_\Lambda\|_\infty$ w.r.t. y, α, μ, while respecting the equality constraints and converting the strict inequality constraint $\mu > 0$ to $\mu \geq \varepsilon\mathbb{1}$ for a small ε, e.g., $\varepsilon = 10^{-8}$, to ensure that the inequality is satisfied strictly. This results in a linear program (LP),

$$\begin{aligned} \min_{t,y,\alpha,\mu} t \qquad \text{s.t.} \quad & -t\mathbb{1} \leq \alpha_\Lambda \leq t\mathbb{1}, \\ & A^\top y = \Omega^\top \alpha + \Psi\mu, \\ & \mu \geq \varepsilon\mathbb{1}\ . \end{aligned}$$

For the optimal solution $(t^*, y^*, \alpha^*, \mu^*)$ we have by definition the smallest possible $t := \|\alpha_\Lambda\|_\infty$. If t^* is not smaller than one, then no y exists with smaller t. We therefore declare $\overline{x}$ the unique minimizer if $t^* < 1$, and if $t^* \geq 1$, $\overline{x}$ cannot be the unique minimizer. Numerically, we would test whether $t^* \leq 1 - \varepsilon$. Technically, by applying the above procedure we risk rejecting a unique solution $\overline{x}$, for which $1 - \varepsilon < t^* < 1$. Recall that the choice for ε is ad hoc. In order to resolve this issue, we provide next a theoretically well-established methodology to deal with the *strict* feasibility problem

$$\begin{aligned} & \Omega^\top \alpha + \Psi\mu = A^\top y, \\ & \alpha_{\Lambda^c} = \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}), \\ & \|\alpha_\Lambda\|_\infty < 1, \\ & \mu > 0. \end{aligned} \tag{3.26}$$

The above feasibility problem is recast as a linear system of inequalities, as we will see next. For this purpose, we transform the above problem into the form $Mz = q, Pz < p$.

Theorem 3.11. *Let M, P be a matrices such that $\mathcal{N}(M^\top) = \{0\}$. Then, there is a point $\overline{z}$ with $M\overline{z} = q, P\overline{z} < d$ if and only if $v = 0$ is the only feasible solution of the problem*

$$q^\top u + p^\top v \leq 0, \quad M^\top u + P^\top v = 0, \quad v \geq 0. \tag{3.27}$$

Proof. We consider the following pair of linear programs

$$\begin{aligned} &\max\ 0 && \text{(P)} && \min\ q^\top u + d^\top v && \text{(DP)} \\ &\text{s.t. } Mz = q && && \text{s.t. } M^\top u + P^\top v = 0 \\ &\quad Pz \leq d, && && \quad v \geq 0 \end{aligned}$$

where (P) is the primal problem and (DP) its dual.

First, note that under the assumption $\mathcal{N}(M^\top) = \{0\}$, vector $v = 0$ is the only feasible solution of (3.27) if and only if $v = 0$ is the only solution of (DP). Indeed, the common constraint $M^\top u + P^\top v = 0$ implies due to the assumptions $v = 0$ and $\mathcal{N}(M^\top) = \{0\}$ that $(u, v) = (0, 0)$.

Hence, we show the statement of the theorem by showing that (P) is strictly feasible if and only if $v = 0$ is the only solution of (DP). On the one hand, we assume that (P) is strictly feasible. Since any feasible solution is also a solution of (P), we deduce from the existence of a primal solution the existence of a dual solution. The optimality conditions yield in particular $(d - P\overline{z})^\top v^* = 0$ for such a primal solution $\overline{z}$ of (P) and any dual solution (u^*, v^*) of (DP). Since $P\overline{z} < d$ holds, it follows that $v^* = 0$ is the only solution. On the other hand, we assume for each dual solution (u^*, v^*) that $v^* = 0$. In view of the strict complementarity condition [Van14, Thm 10.7], it exists a pair of primal and dual solutions such that

$$(d - Pz^*) + v^* > 0. \tag{3.28}$$

Since by assumption $v^* = 0$, (3.28) implies the existence of a primal solution z^* with $Pz^* < d$, which concludes the proof. □

In view of Theorem 3.11, we check feasibility of the system (3.26) by linear programming. First, we transform (3.26) into the primal problem of Theorem 3.11:

$$\begin{aligned} &\max_{\alpha,\mu,y}\ 0 \\ &\text{s.t. } \underbrace{\begin{pmatrix} \Omega^\top{}_{\Lambda,\bullet} & \Psi & A^\top \end{pmatrix}}_{=:M} \underbrace{\begin{pmatrix} \alpha \\ \mu \\ y \end{pmatrix}}_{=:z} = \underbrace{-\Omega^\top_{\Lambda^c,\bullet}\,\mathrm{sign}(D_{\Lambda^c,\bullet}\overline{x})}_{=:q}, \\ &\quad \underbrace{\begin{pmatrix} -I & 0 & 0 \\ I & 0 & 0 \\ 0 & -I & 0 \end{pmatrix}}_{=:P} \underbrace{\begin{pmatrix} \alpha \\ \mu \\ y \end{pmatrix}}_{=:z} \leq \underbrace{\begin{pmatrix} \mathbb{1} \\ \mathbb{1} \\ 0 \end{pmatrix}}_{=:d}. \end{aligned} \tag{P$_<$}$$

Assumption $\mathcal{N}(M^\top) = \{0\}$ in Theorem 3.11 corresponds to the first condition from Theorem 3.3. Consequently, if the condition (i) of Theorem 3.3 is satisfied, then $\overline{x}$ is the unique minimizer if and only if

$$\begin{aligned} \max\ & \langle \mathbb{1}, u^+ + u^- + w \rangle \\ \text{s.t. } & \langle \mathbb{1}, u^+ + u^- \rangle \leq \langle \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}), \Omega_{\Lambda^c,\bullet} v \rangle, \\ & \Omega_{\Lambda,\bullet} v = u^+ - u^-, \\ & \Psi v = w, \\ & Av = 0, \\ & u^+, u^-, w \geq 0, \end{aligned} \tag{DP$_<$}$$

has an optimal objective value equal to zero.

Remark 3.2. Any solution to (DP$_<$), yielding a nonzero objective value, is a counter-example to the nullspace property in Theorem 3.10.

3.2 Probabilistic Recovery

In this section, we investigate the probability that a solution of (3.2) is unique under the assumption that only its co-/sparsity is known. More precisely, we are looking for the likelihood of a unique solution of (3.2) given a fixed number of random linear measurements and the sparsity.

For $f = \|\cdot\|_1$ and a Gaussian matrix[2] A, it is well-known that the recovery of a sparse vector x from exact measurements Ax depends only on its sparsity level, and is independent of the locations or values of the nonzero entries. There are precise relations between the sparsity s, ambient dimension n, and the number of samples m that guarantee success of (3.2a). Moreover, several authors have shown that there is a transition from absolute success to absolute failure, and they have *accurately* characterized in the case of regularization with ℓ_1 or other *norms* the location the location of this transition, also called a *phase transition*. For the TV-seminorm, such a complete analysis is still missing. As mentioned in the introduction, the authors in [Zha16] have recently shown that in the case of one-dimensional TV-minimization, case (3.2d), the phase transition is accurately described by an effective bound for the statistical dimension of a descent cone, which is based on the squared distance of a standard normal vector to the subdifferential of the objective function at the sought solution $\overline{x}$,

$$\min_{\tau \geq 0} \mathbb{E}[\operatorname{dist}(X, \tau \partial f(\overline{x})], \qquad X \sim \mathcal{N}(0, I).$$

Next, we investigate whether the same relation that describes phase transitions for ℓ_1-minimization and 1D TV-minimization also holds for all our objective functions in (3.2a)-(3.2f).

[2] Every entry is normal distributed, i.e., $a_{ij} \sim \mathcal{N}(0,1)$.

3.2.1 Statistical Dimension Estimation

Firstly, we compile recent results of convex signal reconstruction with a Gaussian sampling model and briefly explain how these results lead to a sharp bound for the number of Gaussian measurements that suffice to recover $\overline{x}$ from (3.2) uniquely.

We start by defining the *statistical dimension*, which generalizes the concept of dimension of subspaces to close, convex cones.

Definition 3.12. (Statistical dimension, [Ame14]) The *statistical dimension* $\delta(K)$ of a closed, convex cone $K \subseteq \mathbb{R}^n$ is the quantity

$$\delta(K) := \mathbb{E}[\|\Pi_K(X)\|_2^2], \tag{3.29}$$

where Π_K is the Euclidean projection onto K and X is a standard normal vector, i.e., $X \sim \mathcal{N}(0, I)$.

In the subsequent proposition, we gather some essential properties of the statistical dimension, which also reveal that the statistical dimension generalizes the concept of the dimension of subspaces to closed, convex cones.

Proposition 3.13 ([Ame14, Prop. 3.1])**.** *Let $K \subseteq \mathbb{R}^n$ be a closed, convex cone. The statistical dimension obeys the following laws.*

1. ***Polar Formulation.*** *The statistical dimension is expressed in terms of the* ***polar cone***[3]

$$\delta(K) = \mathbb{E}[\operatorname{dist}(X, K^\circ)^2], \tag{3.30}$$

where K and $X \sim \mathcal{N}(0, I)$.

2. ***Rotational invariance.*** *The statistical dimension does not depend on the orientation of the cone*

$$\delta(UK) = \delta(K), \tag{3.31}$$

for each orthogonal matrix $U \in \mathbb{R}^{n \times n}$.

3. ***Subspaces.*** *For each subspace L, the statistical dimension satisfies*

$$\delta(L) = \dim(L). \tag{3.32}$$

4. ***Complementarity.*** *The sum of the statistical dimension of a cone and that of its polar equals the ambient dimension*

$$\delta(K) + \delta(K^\circ) = n. \tag{3.33}$$

This generalizes the property $\dim(L) + \dim(L^\top) = n$ for each subspace $L \subseteq \mathbb{R}^n$.

[3] $K^\circ = \{y \in \mathbb{R}^n \mid \langle y, x\rangle \leq 0, \forall x \in K\}$

5. ***Monotonicity.*** *For each closed convex cone $K \subseteq \mathbb{R}^n$ the inclusion $C \subseteq K$ implies that $\delta(C) \leq \delta(K)$.*

Another quantity used in the same context in literature is the Gaussian width. For completeness we give the definition and remark the relation to the statistical dimension.

Definition 3.14. (Gaussian width, [Cha12]) The *Gaussian width* of $C \subseteq \mathbb{R}^n$ is

$$\omega(C) := \mathbb{E}[\sup_{z \in C} \langle X, z \rangle],$$

where X is a standard normal vector, i.e., $x_i \sim \mathcal{N}(0,1)$ for $i \in [n]$.

Remark 3.3. In case of closed, convex cones, see [Ame14], we have the following relationship between the *statistical dimension* and the *Gaussian width*:

$$\omega(K)^2 \leq \delta(K) \leq \omega(K)^2 + 1.$$

Before stating the main result of this subsection, we give an alternative characterization of the uniqueness of a solution $\overline{x}$ to (3.2). To this end, we define the *descent cone*.

Definition 3.15. (Descent cone) The *descent cone* of a proper convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ at a point $\overline{x}$ is

$$\mathcal{D}_f(\overline{x}) := \operatorname{cone}(\{z - \overline{x} \colon f(z) \leq f(\overline{x})\}), \tag{3.34}$$

i.e., the conic hull of the directions that do not increase f near $\overline{x}$.

Proposition 3.16 ([FR13, Thm. 4.35])**.** *The vector $\overline{x}$ is the* unique *solution of the convex program* (3.2) *if and only if*

$$\mathcal{D}_f(\overline{x}) \cap \mathcal{N}(A) = \{0\}. \tag{3.35}$$

The next theorem uses the statistical dimension to bound the probability that two convex cones share a ray.

Theorem 3.17. *[Ame14, Thm. I] Fix a tolerance parameter $\varepsilon \in (0,1)$. Let C and K be convex cones in $\mathbb{R}^n$, and draw a random orthogonal matrix $Q \in \mathbb{R}^{n \times n}$. Then*

- $\delta(C) + \delta(K) \leq n - a_\varepsilon \sqrt{n} \implies C \cap QK \neq \{0\}$ *with probability* $\leq \varepsilon$*;*
- $\delta(C) + \delta(K) \geq n + a_\varepsilon \sqrt{n} \implies C \cap QK \neq \{0\}$ *with probability* $\geq 1 - \varepsilon$*,*

with $a_\varepsilon := \sqrt{8 \log(4/\varepsilon)}$.

In the sequel, we consider $A \in \mathbb{R}^{m \times n}$ to be a Gaussian matrix. Then with high probability $\dim(\mathcal{N}(A)) = n - m$. Further, we have that $\mathcal{N}(A)$ is a convex cone and $\delta(\mathcal{N}(A)) = n - m$. Consequently, we use condition (3.35) to derive using Theorem 3.17, the probability that $\overline{x}$ is the unique solution of (3.2).

Theorem 3.18. *[Ame14, Thm. II] Fix a tolerance parameter $\varepsilon \in (0,1)$. Let $\overline{x} \in \mathbb{R}^n$ be a fixed vector, and let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper convex function. Suppose $A \in \mathbb{R}^{m\times n}$ has independent standard normal entries, and let $b = A\overline{x}$. Then*

- $m \leq \delta(\mathcal{D}_f(\overline{x})) - a_\varepsilon\sqrt{n} \implies$ (3.2) *succeeds with probability* $\leq \varepsilon$;
- $m \geq \delta(\mathcal{D}_f(\overline{x})) + a_\varepsilon\sqrt{n} \implies$ (3.2) *succeeds with probability* $\geq 1-\varepsilon$,

with $a_\varepsilon := \sqrt{8\log(4/\varepsilon)}$.

Theorem 3.18 states that there is a transition from failure to success and localizes the phase transition at $m = \delta(\mathcal{D}_f(\overline{x}))$. Consequently, the statistical dimension of the descent cone bounds the needed number of linear measurements to guarantee exact recovery by (3.2) [Ame14].

3.2.2 Upper Bounds for the Statistical Dimension

As aforementioned, the calculation of the statistical dimension might be difficult. In the following, we discuss an upper bound that admits a closed form in some cases.

By connecting the subdifferential to the descent cone, we provide an alternative formulation of the statistical dimension.

Proposition 3.19 ([Tro15, Fact 4.4])**.** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper convex function and $x \in \mathbb{R}^n$. If the subdifferential $\partial f(x)$ is nonempty and does not contain the origin, then*

$$\mathcal{D}(f,x)^\circ = \overline{\operatorname{cone}}(\partial f(x)). \tag{3.36}$$

Combining the proposition above and (3.30), we calculate the statistical dimension in terms of the subdifferential by

$$\begin{aligned}\delta(\mathcal{D}(f,x)) &= \mathbb{E}[\operatorname{dist}(X,\mathcal{D}(f,x)^\circ)^2]\\ &= \mathbb{E}[\operatorname{dist}(X,\overline{\operatorname{cone}}(\partial f(x)))^2]\\ &= \mathbb{E}[\inf_{\tau\geq 0}\operatorname{dist}(X,\tau\cdot\partial f(x))^2].\end{aligned}$$

Next, we introduce the upper bound from [Ame14], which uses the connection between the descent cone and the subdifferential.

Proposition 3.20 ([Ame14, Prop. 4.1])**.** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper convex function, and let $x \in \mathbb{R}^n$. Assuming that the subdifferential $\partial f(x)$ is non-empty, compact, and does not contain the origin. Define the function*

$$J(\tau) := J(\tau;\partial f(x)) := \mathbb{E}\left[\operatorname{dist}\left(g,\tau\cdot\partial f(x)\right)^2\right], \qquad \text{for } \tau \geq 0, \tag{3.37}$$

where g is a normally distributed, i.e., $g_i \sim \mathcal{N}(0,1)$ for $i \in [n]$. Then, we have the upper bound

$$\delta(\mathcal{D}(f,x)) \leq \inf_{\tau\geq 0} J(\tau). \tag{3.38}$$

Furthermore, the function J is strictly convex, continuous at $\tau = 0$, and differentiable for $\tau \geq 0$. It achieves its minimum at a unique point.

As an assumption in Theorem 3.20, the subdifferential needs to be compact. Considering the functions (3.2a)-(3.2f) from the introduction, this assumption might be violated. Indeed, let $x \in \{0,1\}^n$. The subdifferential of the indicator function $\delta_{[0,1]^n}$ is written as

$$\partial\delta_{[0,1]^n}(\overline{x}) = \{\Psi y \mid y \geq 0\}, \tag{3.39}$$

where Ψ is the matrix from Definition 3.2. Consequently, $\partial\delta_{[0,1]^n}(\overline{x})$ is unbounded and hence not compact. Hence, the subdifferentials of the functions (3.2b),(3.2c),(3.2e) and (3.2f) are not compact. As a result, Proposition 3.20 is not directly applicable in presence of box constraints.

Inspecting the proof of Proposition 3.20 given in [Ame14, Prop. 4.1], the compactness is not used to prove the upper bound (3.38) but used for showing the properties of J defined in (3.37). As a remedy, we next prove that even if the subdifferential is not compact, the function (3.37) is still

(a) strictly convex,

(b) continuous at $\tau = 0$,

(c) differentiable for $\tau \geq 0$,

(d) achieves its minimum at a unique point.

Before proving the properties of J from (3.37), in case of a non-compact subdifferential, we need the following lemma.

Lemma 3.21. *Consider $S, B, C \subseteq \mathbb{R}^n$ non-empty, convex and closed, with $S = B + C$. If B is compact, then for every $\overline{x} \in \mathbb{R}^n$ it exists $p \in \mathbb{R}^n$ such that*

$$\pi_S(\overline{x}) = \pi_C(\overline{x}) + p \tag{3.40}$$

where $\|p\|_2 \leq 2 \cdot \sup_{x \in B} \|x\|_2$.

Proof. Since S is closed and convex, we choose

$$(c^*, b^*) \in \arg\min_{\substack{b \in B \\ c \in C}} \|b + c - \overline{x}\|_2$$

to obtain

$$\pi_S(\overline{x}) = \underbrace{\arg\min_{c \in C} \|\overline{x} - c - b^*\|_2}_{=\pi_C(\overline{x} - b^*)} + \underbrace{\arg\min_{b \in B} \|\overline{x} - c^* - b\|_2}_{=\pi_B(\overline{x} - c^*)}.$$

By a zero expansion, we formulate $\pi_S(\overline{x})$ using the results from above as

$$\begin{aligned}\pi_S(\overline{x}) &= \pi_C(\overline{x}) + \underbrace{\pi_B(\overline{x} - c^*) + \pi_C(\overline{x} - b^*) - \pi_C(\overline{x})}_{=:p} \\ &= \pi_C(\overline{x}) + p.\end{aligned}$$

Considering that the projection on a closed convex set is a contraction and that B is bounded, we obtain

$$\|\pi_C(\overline{x} - b^*) - \pi_C(\overline{x})\|_2 \leq \|\overline{x} - b^* - \overline{x}\|_2 = \|b^*\|_2 \leq \sup_{x \in B} \|x\|_2 < \infty.$$

Finally, we calculate an upper bound to p by

$$\begin{aligned}\|p\|_2 &\leq \|\pi_B(\overline{x} - c^*)\|_2 + \|\pi_C(\overline{x} - b^*) - \pi_C(\overline{x})\|_2 \\ &\leq 2 \cdot \sup_{x \in B} \|x\|_2.\end{aligned}$$

□

Before verifying the desired properties for function J from (3.37), we validate them for a slightly simpler function

$$\mathcal{J}_u(\tau) := \operatorname{dist}\left(u, \tau \cdot \partial f(x)\right)^2. \tag{3.41}$$

Lemma 3.22. *The function $\mathcal{J}_u$ from (3.41) is convex in τ.*

Proof. Setting $S := \partial f(x)$, we rewrite $\mathcal{J}_u(\tau)$ for $\tau > 0$ to

$$\mathcal{J}_u(\tau) = \left(\inf_{s \in S} \|u - \tau s\|_2\right)^2 = \left(\tau \inf_{s \in S} \left\|\frac{u}{\tau} - s\right\|_2\right)^2 = \left(\tau \operatorname{dist}\left(\frac{u}{\tau}, S\right)\right)^2. \tag{3.42}$$

Due to the convexity of the distance function [Roc97, p. 34], the perspective transformation of a convex function is convex [HL93a, p. 160], and the square of a non-negative convex function is convex. Hence $\mathcal{J}_u(\tau)$ is convex. □

Lemma 3.23. *Let $B, C \subseteq \mathbb{R}^n$ be closed, convex sets such that $\partial f(x) = B + C$, B is compact and C is a cone. Then the function $\mathcal{J}_u$ from (3.41) is continuous at $\tau = 0$ and $\mathcal{J}_u(0) = \|\pi_C(u) - u\|_2^2$.*

Proof. First, we observe that for any $\varepsilon > 0$ it follows that $\varepsilon(B + C) = \varepsilon B + C$. We now prove continuity at $\tau = 0$ by showing that for any sequence $\varepsilon \to 0$ it follows that $\mathcal{J}_u(\varepsilon) \to \mathcal{J}_u(0) = \|\pi_C(u) - u\|_2^2$. We remind that $\operatorname{dist}\left(u, \tau\partial f(x)\right)^2$ represents

$$\min_{\substack{b \in B \\ c \in C}} \|u - \tau \cdot (b + c)\|_2^2 \tag{3.43}$$

suggesting the assumption $\mathcal{J}_u(0) = \|\pi_C(u) - u\|_2^2$.

Therefore, we calculate

$$\begin{aligned}
|\mathcal{J}_u(0) - \mathcal{J}_u(\varepsilon)| &= \left| \|\pi_C(u) - u\|_2^2 - \|\pi_{C+\varepsilon B}(u) - u\|_2^2 \right| \\
&\overset{①}{=} \left| \|\pi_C(u) - u\|_2^2 - \|\pi_C(u) + p - u\|_2^2 \right| \\
&= \left| \|\pi_C(u)\|_2^2 - \|\pi_C(u) + p\|_2^2 + 2\langle u, p\rangle \right| \\
&\leq \left| \|\pi_C(u)\|_2^2 - \|\pi_C(u) + p\|_2^2 \right| + 2\left|\langle u, p\rangle\right| \\
&\overset{②}{\leq} \|p\|_2^2 + 2\left|\langle u, p\rangle\right| \\
&\leq (2\varepsilon\overline{B})^2 + 2\|u\|_2(2\varepsilon\overline{B}) \to 0 \text{ if } \varepsilon \to 0,
\end{aligned}$$

where $\overline{B} := \sup_{x\in B} \|x\|_2^2$ and ①, ② follows by Lemma 3.21. □

Lemma 3.24. *Let $B, C \subseteq \mathbb{R}^n$ closed, convex sets such that $\partial f(x) = B + C$ with B a compact set and C a cone. Assume that the subdifferential $\partial f(x)$ does not contain the origin. Then the function* (3.41) *attains its minimum in the interval*

$$\left[0, \frac{\|u\|_2 + \|u - \pi_C(u)\|_2}{s}\right], \tag{3.44}$$

where $s \leq \|y\|_2$ for all $y \in \partial f(x)$.

Proof. We assume that $s \leq \|y\|_2$ for all $y \in \partial f(x)$ and $\tau \geq \frac{\|u\|_2}{s}$. Then we calculate

$$\begin{aligned}
\sqrt{\mathcal{J}_u(\tau)} &= \operatorname{dist}(u, \tau\partial f(x)) \\
&= \inf_{y\in\partial f(x)} \|\tau y - u\|_2 \\
&\geq \inf_{y\in\partial f(x)} \tau\|y\|_2 - \|u\|_2 \\
&\geq \tau s - \|u\|_2 \geq 0.
\end{aligned}$$

By choosing $\tau > \frac{\|u\|_2 + \|u - \pi_C(u)\|_2}{s}$, we obtain

$$\|u - \pi_C(u)\|_2^2 = \mathcal{J}_u(0) < \mathcal{J}_u(\tau).$$

Given the convexity of $\mathcal{J}_u$ and just proven boundedness, it follows that the minimum of $\mathcal{J}_u$ is attained in the interval (3.44). □

Lemma 3.25. *The function* (3.41) *is continuously differentiable on $(0, \infty)$ and the derivative is given by*

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{J}_u(\tau) = -\frac{2}{\tau}\langle \pi_{\tau S}(u), u - \pi_{\tau S}(u)\rangle, \tag{3.45}$$

where $S := \partial f(x)$ is a closed convex set. Moreover, the right derivative at $\tau = 0$ exists.

Proof. The squared distance $\operatorname{dist}^2(x, \mathcal{X})$ w.r.t. a closed convex set $\mathcal{X} \subseteq \mathbb{R}^n$ is a continuously differentiable function for all $x \in \mathbb{R}^n$, see [Ame14, Sec. B.2]. Applying [RW10,

Thm. 2.26], we obtain the derivative

$$\nabla \operatorname{dist}^2(x, \mathcal{X}) = 2(x - \pi_{\mathcal{X}}(x)). \tag{3.46}$$

Rewriting $\mathcal{J}_u$ as in (3.42), we are able to use (3.46) to obtain

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{J}_u(\tau) = \frac{\mathrm{d}}{\mathrm{d}\tau}\tau^2 \operatorname{dist}\left(\frac{u}{\tau}, S\right)^2 \tag{3.47}$$

$$= \frac{2}{\tau}\left(\operatorname{dist}(u, \tau S)^2 - \langle u, u - \pi_{\tau S}(u)\rangle\right), \tag{3.48}$$

where $\tau > 0$.

Reminding that $\operatorname{dist}(u, \tau S)^2 = \|u - \pi_{\tau S}(u)\|_2^2$, we obtain from (3.48) the desired derivative (3.45).

So far, we calculated that $\mathcal{J}_u$ is continuously differentiable on $(0, \infty)$. The result [Roc97, Thm. 24.1] implies that if a convex function is continuously differentiable on $(0, \infty)$ and continuous on $[0, \infty]$, the right derivative at zero exists. The function $\mathcal{J}_u$ is a convex by Lemma 3.22 and by Lemma 3.23 continuous on $[0, \infty]$. Thus, it follows that the right derivative

$$\mathcal{J}_u'(0) = \lim_{\tau \downarrow 0} \frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{J}_u(\tau) \tag{3.49}$$

does exist. □

Lemma 3.26. *Let $B, C \subseteq \mathbb{R}^n$ closed, convex sets such that $\partial f(x) = B + C$ with B a compact set and C a cone. Then the derivative of (3.41) is bounded through*

$$|\mathcal{J}_u'(\tau)| \leq 4\overline{B}\left(\|u\|_2 + 2\|\pi_C(u)\|_2 + 2\tau\overline{B}\right) \tag{3.50}$$

where $\overline{B} := \sup_{x \in B} \|x\|_2$.

Proof. Observing that for any $\varepsilon > 0$ it follows that $\varepsilon(B + C) = \varepsilon B + C$, we calculate

$$\begin{aligned}
|\mathcal{J}_u'(\tau)| &= \frac{2}{\tau}\left|\langle \pi_{\tau S}(u), u - \pi_{\tau S}(u)\rangle\right| \\
&\overset{①}{=} \frac{2}{\tau}\left|\langle \pi_C(u) + p, u - \pi_C(u) - p\rangle\right| \\
&\overset{②}{=} \frac{2}{\tau}\left|\langle \pi_C(u), -p\rangle + \langle p, u - \pi_C(u) - p\rangle\right| \\
&= \frac{2}{\tau}\left|\langle p, u - 2\pi_C(u) - p\rangle\right| \\
&\leq \frac{2}{\tau}\|p\|_2\|u - 2\pi_C(u) - p\|_2 \\
&\overset{③}{\leq} 4\overline{B}\left(\|u\|_2 + 2\|\pi_C(u)\|_2 + 2\tau\overline{B}\right).
\end{aligned}$$

In ① and ③, we use Lemma 3.21 with $\overline{B} := \sup_{x \in B} \|x\|_2$. In ②, we use that $\langle \pi_C(u), u - \pi_C(u)\rangle = 0$ since C is a closed convex cone. □

Finally, we extend Proposition 3.20 in case when $\partial f(x)$ is not compact.

Proposition 3.27. *Let $C, B \subseteq \mathbb{R}^n$ be closed convex sets, such that $\partial f(x) = C + B$. Assume that $\partial f(x)$ does not contain the origin, B is compact and C is a cone. Then (3.37) is strictly convex, continuous at $\tau = 0$ and differentiable for $\tau > 0$. Furthermore, it attains its minimum at a unique point and*

$$J'(\tau) = \mathbb{E}\left[\mathcal{J}_u'(\tau)\right] \text{ for all } \tau \geq 0, \tag{3.51}$$

where u is normal random vector, i.e., $u \sim \mathcal{N}(0,1)$ for $i \in [n]$.

Proof. Further in this proof, we need an upper bound to $\mathbb{E}[\|u\|_2]$ which we obtain by using Jensen's inequality

$$\begin{aligned} & \mathbb{E}[\|u\|_2]^2 \leq \mathbb{E}[\|u\|_2^2] = n \\ \Leftrightarrow\ & \mathbb{E}[\|u\|_2] \leq \sqrt{n}. \end{aligned}$$

Since $\mathbb{E}[\|\pi_C(u)\|_2] \leq \mathbb{E}[\|u\|_2]$, it directly follows that $\mathbb{E}[\|\pi_C(u)\|_2] \leq \sqrt{n}$.

Continuity at $\tau = 0$ follows directly by using Lemma 3.23 and $\mathbb{E}[\|u\|_2] \leq \sqrt{n}$ in view of

$$\begin{aligned} J(\varepsilon) - J(0) =\ & \mathbb{E}\left[\mathcal{J}_u(\varepsilon) - \mathcal{J}_u(0)\right] \\ & \leq (3\varepsilon\overline{B})^2 + 2\sqrt{n}(3\varepsilon\overline{B}) \to 0 \text{ if } \varepsilon \to 0. \end{aligned}$$

Differentiability follows by applying *differentiation under the integral sign* [Els18, Satz 5.7]. This theorem tells us, if it exists a function $g : \mathbb{R}^n \to \mathbb{R}$ so that it holds for any relative open interval $I \subset \mathbb{R}_{\geq 0}$:

$$\left|\mathcal{J}_u'(\tau)\right| \leq g(u) \qquad \forall u \in \mathbb{R}^n, \tau \in I,$$

with $\mathbb{E}[g(u)] < \infty$. Then, we may write $J'(\tau) = \frac{\mathrm{d}}{\mathrm{d}\tau}\mathbb{E}\left[\mathcal{J}_u(\tau)\right] = \mathbb{E}\left[\mathcal{J}_u'(\tau)\right]$. Lemma 3.25 yields such a function g by

$$\begin{aligned} g(u) &= 4\overline{B}\left(\|u\|_2 + 2\|\pi_C(u)\|_2 + 2\overline{B}\sup_{\tau' \in I}\tau'\right) \\ \text{and } \mathbb{E}[g(u)] &\leq 4\overline{B}\left(\sqrt{n} + 2\sqrt{n} + \overline{B}\sup_{\tau' \in I}\tau'\right) < \infty. \end{aligned}$$

Next, we show that J is *strictly convex*. Since $\mathcal{J}_u$ is convex, it follows that for every $0 \leq \tau < \rho$ and $\eta \in (0,1)$ it holds

$$\begin{aligned} \mathbb{E}\left[\mathcal{J}_u(\eta\rho + (1-\eta)\tau)\right] \leq\ & \mathbb{E}\left[\eta \cdot \mathcal{J}_u(\rho) + (1-\eta)\cdot\mathcal{J}_u(\tau)\right] \\ & = \eta \cdot \mathbb{E}\left[\mathcal{J}_u(\rho)\right] + (1-\eta)\,\mathbb{E}\left[\cdot\mathcal{J}_u(\tau)\right], \end{aligned}$$

implying that J is convex. We prove that J is strictly convex by contradiction. If J is not strictly convex, there have to be $0 \leq \rho < \tau$ and $\eta \in (0,1)$ so that the above

inequality holds with equality. Equality would imply that

$$\mathcal{J}_u(\eta\rho + (1-\eta)\tau) = \eta \cdot \mathcal{J}_u(\rho) + (1-\eta) \cdot \mathcal{J}_u(\tau), \tag{3.52}$$

for almost every $u \in \mathbb{R}^n$. Setting $u = 0$, we calculate

$$\begin{aligned}\mathcal{J}_0(\eta\rho + (1-\eta)\tau) &= \operatorname{dist}^2(0, (\eta\rho + (1-\eta)\tau)\partial f(x)) = (\eta\rho + (1-\eta)\tau)^2 \inf_{s\in\partial f(x)} \|s\|_2^2 \\ &< (\eta\rho^2 + (1-\eta)\tau^2) \inf_{s\in\partial f(x)} \|s\|_2^2 \\ &= \eta\mathcal{J}_0(\rho) + (1-\eta)\mathcal{J}_0(\tau).\end{aligned}$$

Since $\mathcal{J}_u(\tau)$ is continuous in u, it exists $\varepsilon > 0$ such that

$$\mathcal{J}_0(\eta\rho + (1-\eta)\tau) < \eta\mathcal{J}_0(\rho) + (1-\eta)\mathcal{J}_0(\tau),$$

for all $\|u\| < \varepsilon$. This is a contradiction to (3.52), hence J is strictly convex.

Finally, we prove *the minimum of J is attained at a unique point.* Therefore, we set $s := \inf_{y\in\partial f(x)} \|y\|_2$ and assume that $\tau \geq \frac{\sqrt{n}}{s}$. This gives

$$\begin{aligned}J(\tau) &\overset{①}{\geq} \mathbb{E}\left[\mathcal{J}_u(\tau) \;\middle|\; \|u\|_2 \leq \sqrt{n}\right] \cdot \mathbb{P}\left(\|u\|_2 \leq \sqrt{n}\right) \\ &\overset{②}{\geq} \frac{1}{2}\,\mathbb{E}\left[\mathcal{J}_u(\tau) \;\middle|\; \|u\|_2 \leq \sqrt{n}\right] \\ &\overset{③}{\geq} \frac{1}{2}\,\mathbb{E}\left[(\tau s - \|u\|_2)^2 \;\middle|\; \|u\|_2 \leq \sqrt{n}\right] \\ &\overset{④}{\geq} \frac{1}{2}\,\mathbb{E}\left[\tau s - \|u\|_2 \;\middle|\; \|u\|_2 \leq \sqrt{n}\right]^2 \\ &\overset{⑤}{\geq} \frac{1}{2}\left(\tau s - \sqrt{n}\right)^2.\end{aligned}$$

The properties used to achieve the inequalities above are

① law of total expectation;
② $\mathbb{E}[\|u\|_2] \leq \sqrt{n}$ implies $\mathbb{P}(\|u\|_2 \leq \sqrt{n}) \geq \frac{1}{2}$;
③ the bound from Lemma 3.24;
④ Jensen's inequality;
⑤ the bound $\mathbb{E}[\|u\|_2] \leq \sqrt{n}$.

Since C is a closed convex cone we have $J(0) = \mathbb{E}\left[\|\pi_C(u) - u\|_2^2\right] = \delta(C^\circ) \leq n$. Consequently, $\tau > \frac{2\sqrt{n}}{s}$ implies $J(\tau) > J(0)$. Using the strict convexity of J, we showed that the minimum is attained at a unique point in the interval $\left[0, \frac{2\sqrt{n}}{s}\right]$. □

Corollary 3.28. *Let $C, B \subseteq \mathbb{R}^n$ closed convex sets, such that $\partial f(x) = C + B$. Assume that $\partial f(x)$ does not contain the origin, B is compact and C is a cone. Then*

$$\delta(\mathcal{D}(f, x)) \leq \delta(C^\circ). \tag{3.53}$$

Proof. Proposition 3.20 provides the bound

$$\delta(\mathcal{D}(f,x)) \leq \inf_{\tau\geq 0} J(\tau). \tag{3.54}$$

By the proof of Proposition 3.27 we have

$$\inf_{\tau\geq 0} J(\tau) \leq J(0) \leq \delta(C^\circ). \tag{3.55}$$

□

Error Bounds There is a useful tool for bounding the statistical dimension in terms of $\partial f(\overline{x})$. Interestingly, this bound is tight for some classes of f, e.g., norms.

Below, we present two error bounds for the bound of Proposition 3.20 estimating the distance between the upper bound and the statistical dimension.

Theorem 3.29. *[Ame14, Thm. 4.3.] Let $f : \mathbb{R}^n \to \mathbb{R}$ be a norm on $\mathbb{R}^n$ and let $x \in \mathbb{R}^n \setminus \{0\}$. Then*

$$0 \leq \inf_{\tau\geq 0} J(\tau) - \delta(\mathcal{D}_f(\overline{x})) \leq \frac{2\sup\{\|p\| : p \in \partial f(\overline{x})\}}{f(\overline{x}/\|\overline{x}\|)}. \tag{3.56}$$

For the ℓ_1-case the right-hand side can be made very accurate for large sparsity parameters s providing an accurate estimate of the statistical dimension of the ℓ_1-descent cone. On the other hand, this error estimate is not very accurate when the sparsity s is small. The work in [FM14] contains a bound that improves this result and also extends to more general functions.

Theorem 3.30. *[FM14, Prop. 1] Suppose that, for $\overline{x} \neq 0$, the subdifferential $\partial f(\overline{x})$ is weakly decomposable, i.e.,*

$$\exists \overline{p} \in \partial f(\overline{x}) \quad \textit{such that} \quad \langle p - \overline{p}, \overline{p}\rangle = 0, \qquad \forall p \in \partial f(\overline{x}). \tag{3.57}$$

Then

$$\delta(\mathcal{D}_f(\overline{x})) \leq \inf_{\tau\geq 0} J(\tau) \leq \delta(\mathcal{D}_f(\overline{x})) + 6, \tag{3.58}$$

with $J(\tau) = \mathbb{E}[\operatorname{dist}(X, \tau\partial f(\overline{x})]$.

This result implies that the upper bound is almost tight, when considering (3.2a). Furthermore, the authors of [Zha16] showed that if Ω is the one-dimensional discrete gradient for (3.2d), then the subdifferential is for every $\overline{x}$ weakly decomposable. Hence, the upper bound to the statistical dimension precisely describe the phase transition. Unfortunately, this is not the case if Ω is the two-dimensional discrete gradient as the following counter-example illustrates. To this end, we consider $f(x) = \|\Omega x\|_1$. Then the subgradient at $\overline{x} \in \mathbb{R}^n$ of f is given by

$$\partial f(\overline{x}) = \{\Omega\alpha \mid \alpha_{\Lambda^c} = \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}), \|\alpha_\Lambda\|_\infty \leq 1\}, \tag{3.59}$$

where $\Lambda = \operatorname{cosupp}(\Omega\overline{x})$. For the computation of (3.59), we refer to the Subdifferential Calculus in Section 2.1.1.

Next, we rewrite the condition (3.57) according to our choice of f yielding

$$\exists w \in \mathcal{V} : \langle \Omega^\top v - \Omega^\top w, \Omega^\top w \rangle = 0, \qquad \forall v \in \mathcal{V}, \tag{3.60}$$

where $\mathcal{V} := \{v \mid v_{\Lambda^c} = \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}), \|v_\Lambda\|_\infty \leq 1\}$.

For an arbitrary choice of $w, v \in \mathcal{V}$, we calculate

$$\begin{aligned}
& w^\top \Omega\Omega^\top v = w^\top \Omega\Omega^\top w && (3.61)\\
\Leftrightarrow\ & w^\top (\Omega\Omega^\top)_{\bullet,\Lambda} v_\Lambda + w^\top (\Omega\Omega^\top)_{\bullet,\Lambda^c} v_{\Lambda^c} = w^\top (\Omega\Omega^\top)_{\bullet,\Lambda} w_\Lambda + w^\top (\Omega\Omega^\top)_{\bullet,\Lambda^c} w_{\Lambda^c} && (3.62)\\
\Leftrightarrow\ & w^\top (\Omega\Omega^\top)_{\bullet,\Lambda} v_\Lambda = w^\top (\Omega\Omega^\top)_{\bullet,\Lambda} w_\Lambda, && (3.63)
\end{aligned}$$

where (3.63) results from $w_{\Lambda^c} = v_{\Lambda^c}$. Since (3.63) must hold for every $v \in \mathcal{V}$, we directly obtain

$$\begin{aligned}
& w^\top (\Omega\Omega^\top)_{\bullet,\Lambda} v_\Lambda = 0 \qquad \forall v_\Lambda : \|v_\Lambda\|_\infty \leq 1 && (3.64)\\
\Leftrightarrow\ & (\Omega\Omega^\top)_{\Lambda,\bullet} w = 0 && (3.65)\\
\Leftrightarrow\ & (\Omega\Omega^\top)_{\Lambda,\Lambda} w_\Lambda + (\Omega\Omega^\top)_{\Lambda,\Lambda^c} w_{\Lambda^c} = 0 && (3.66)\\
\Leftrightarrow\ & (\Omega\Omega^\top)_{\Lambda,\Lambda^c} \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}) = -(\Omega\Omega^\top)_{\Lambda,\Lambda} w_\Lambda. && (3.67)
\end{aligned}$$

As a consequence, the subdifferential $\partial f(\overline{x})$ is weakly decomposable if and only if it exists $w \in \mathcal{V}$ satisfying (3.67).

To provide a counter-example showing that the two-dimensional discrete gradient is not weakly decomposable for every choice of $\overline{x}$, we set

$$\overline{x} := \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}^\top. \tag{3.68}$$

In view of (3.67), the matrix $(\Omega\Omega^\top)_{\Lambda,\Lambda}$ is invertible w.r.t. $\overline{x}$. Then, one directly computes

$$w_\Lambda = -(\Omega\Omega^\top)^{-1}_{\Lambda,\Lambda} (\Omega\Omega^\top)_{\Lambda,\Lambda^c} \operatorname{sign}(\Omega_{\Lambda^c,\bullet}\overline{x}) \tag{3.69}$$

and obtains $\|w_\Lambda\|_\infty > 1$. Since $w \in \mathcal{V}$, it must hold $\|w_\Lambda\|_\infty \leq 1$. Hence, $\partial f(\overline{x})$ is not weakly decomposable.

3.2.3 Explicit Bounds for the Statistical Dimension

In this section, we explicitly calculate the bounds from Proposition 3.20 in the cases (3.2a) - (3.2c) and discuss how we approximate the statistical dimension in general.

Starting (3.2a), we consider $\partial f(\overline{x}) = \partial\|\overline{x}\|_1$ with $\overline{x} \in \mathbb{R}^n$ and $S = \operatorname{supp}(\overline{x})$. Then,

we calculate

$$\begin{aligned} J_1(\tau) &= \mathbb{E}\left[\operatorname{dist}^2\left(u, \tau\partial\|x\|_1\right)\right] && (3.70)\\ &= \mathbb{E}\left[\|u_S - \tau\operatorname{sign}(x_S)\|_2^2 + \min_{\|v\|_\infty\le 1}\|u_{S^c} - \tau v\|_2^2\right] && (3.71)\\ &= |S|(1+\tau^2) + |S^c|\,\mathbb{E}\left[(\overline{u} - \operatorname{sign}(\overline{u})\tau)^2 \;\middle|\; |\overline{u}| > \tau\right]\mathbb{P}(|\overline{u}| > \tau) && (3.72)\\ &= |S|(1+\tau^2) + 2|S^c|\,\mathbb{E}\left[(\overline{u} - \tau)^2 \;\middle|\; \overline{u} > \tau\right]\mathbb{P}(\overline{u} > \tau) && (3.73)\end{aligned}$$

where $u, \overline{u}$ are normal random vectors, i.e., $\overline{u}_i, u_i \sim \mathcal{N}(0,1)$ for $i \in [n]$.

Next, we consider (3.2b) implying $\partial f(\overline{x}) = \partial\|\overline{x}\|_1 + \partial\delta_{\mathbb{R}^n_{\ge 0}}(\overline{x})$ with $\overline{x} \in \mathbb{R}^n_{\ge 0}$ and $S = \operatorname{supp}(\overline{x})$. Then, we calculate

$$\begin{aligned} J_2(\tau) &= \mathbb{E}\left[\operatorname{dist}^2\left(u, \tau\partial\|x\|_1 + \partial\delta_{\mathbb{R}^n_{\ge 0}}(\overline{x})\right)\right] && (3.74)\\ &= \mathbb{E}\left[\|u_S - \tau\operatorname{sign}(x_S)\|_2^2 + \min_{v\le\tau}\|u_{S^c} - v\|_2^2\right] && (3.75)\\ &= |S|(1+\tau^2) + |S^c|\,\mathbb{E}\left[(\overline{u} - \tau)^2 \;\middle|\; \overline{u} > \tau\right]\mathbb{P}(\overline{u} > \tau), && (3.76)\end{aligned}$$

where $u, \overline{u}$ are normal random vectors, i.e., $\overline{u}_i, u_i \sim \mathcal{N}(0,1)$ for $i \in [n]$.

Finally, the definition of f in (3.2c) implies $\partial f(\overline{x}) = \partial\|\overline{x}\|_1 + \partial\delta_{[0,1]^n}(\overline{x})$ with $\overline{x} \in \{0,1\}^n$ and $S = \operatorname{supp}(\overline{x})$. Then, we calculate

$$\begin{aligned} J_3(\tau) &= \mathbb{E}\left[\operatorname{dist}^2\left(u, \tau\partial\|x\|_1 + \partial\delta_{[0,1]^n}(\overline{x})\right)\right] && (3.77)\\ &= \mathbb{E}\left[\sum_{i\in S}\min_{\mu\ge\tau}(u_i - \mu)^2 + \sum_{i\in S^c}\min_{\mu\le\tau}(u_i - \mu)^2\right] && (3.78)\\ &= |S|\,\mathbb{E}[(\overline{u} - \tau)^2 \mid \overline{u} < \tau]\,\mathbb{P}(\overline{u} < \tau) + |S^c|\,\mathbb{E}[(\overline{u} - \tau)^2 \mid \overline{u} > \tau]\,\mathbb{P}(\overline{u} > \tau), && (3.79)\end{aligned}$$

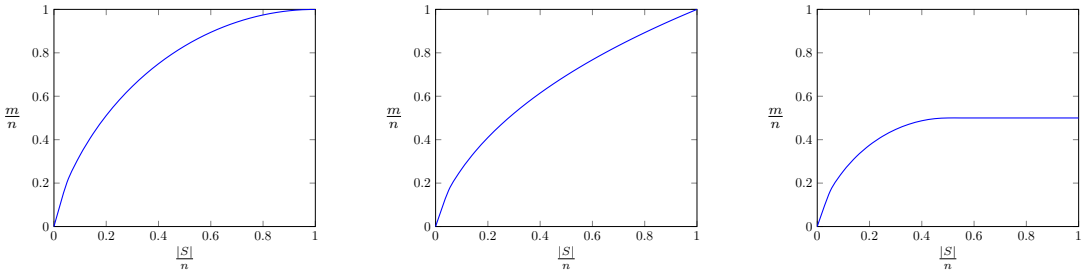
where $u, \overline{u}$ are normal random vectors, i.e., $\overline{u}_i, u_i \sim \mathcal{N}(0,1)$ for $i \in [n]$. We observe that $J_3(0) = \frac{n}{2}$ which coincides with the result of Corollary 3.28.

For better visualization, we display in Figure 3.1 the bound (3.38) with respect to (3.73),(3.76), and (3.79).

Explicit curves for $\min_\tau J(\tau)$ are only computable in the case of (3.2a) - (3.2c). We skip the details here and illustrate these curves in Section 3.3. For (3.2d) - (3.2f) explicit curves for J and for $\min_\tau J(\tau)$ are missing. We use an approximation to the upper bound $J(\tau) = \mathbb{E}\left[\operatorname{dist}^2(u, \tau\partial f(\overline{x}))\right]$ where u is a normal random vector, i.e., $u_i \sim \mathcal{N}(0,1)$ for $i \in [n]$. To approximate the expected value, we use a very large $k \in \mathbb{N}$ and calculate

$$h_k(\tau) = \frac{1}{k}\sum_{i=1}^{k}\operatorname{dist}^2(u_i, \tau\partial f(\overline{x})) =: J_{\text{approx}}(\tau) \approx J(\tau),$$

where each vector u_i is component-wise sampled from the normal distribution $\mathcal{N}(0,1)$.



(a) J_1 (b) J_2 (c) J_3

Figure 3.1: From left to right, we plotted the bound (3.38) with respect to (3.73),(3.76), and (3.79). In particular (c) shows that for binary signals $\frac{n}{2}$ measurements are sufficient for unique recovery.

Note that for computing each $\text{dist}^2(u_i, \tau\partial f(\overline{x}))$ one is faced with solving a quadratic program, i.e.

$$\min \|u_i - \tau y\|_2^2 \qquad \text{subject to} \quad y \in \partial f(\overline{x}).$$

Note that the constraints are linear since $\partial f(\overline{x})$ is described by linear equalities and inequalities.

3.3 Phase Transition Curves

Our objective in this section is to empirically verify whether phase transitions occur at

$$\min_{\tau \geq 0} J(\tau), \tag{3.80}$$

which is a guaranteed tight upper bound to the statistical dimension $\delta(\mathcal{D}(f, \overline{x}))$ if f is (3.2a) or (3.2d) where Ω is the one-dimensional discrete gradient. For the other cases (3.2c), (3.2b), (3.2f) and (3.2e) we provide empirical evidence implying that (3.80) is a tight upper bound to the statistical dimension. We further note, that precise expressions for J are only available for (3.2a) - (3.2c) as shown in the last section. For the cases (3.2d) - (3.2f) using Ω as one- or two-dimensional discrete gradient we construct phase diagrams by numeric recovery rates estimating $\min_{\tau \geq 0} J(\tau)$ and comparing the resulting bounds to the average-case for both solving the reconstruction problems (3.2) and certifying uniqueness by the procedure derived in Section 3.1.3.

3.3.1 1D Empirical Phase Transitions

In this section, we address ℓ_1-minimization and one-dimensional TV-minimization without and with constraints. We define for (3.2d) - (3.2f) the one-dimensional discrete

gradient by

$$\Omega x = \begin{pmatrix} x_1 - x_0 \\ \vdots \\ x_{n-1} - x_{n-2} \end{pmatrix}.$$

Then, we refer to (3.2d) - (3.2f) as one-dimensional TV-minimization.

Testset Our testset consists of several randomly generated signals with specified relative sparsity ρ ranging from 0.05 to 0.95 with step size 0.05. For each of these relative sparsities we create a signal so that

$$\rho \approx \frac{\|\overline{x}\|_0}{n} \text{ resp. } \rho \approx \frac{\|\Omega\overline{x}\|_0}{n}.$$

Hence, we cover almost the full range from highly sparse to dense signals. In addition, we create three different classes of signals: with real-valued, non-negative or binary entries, i.e., $x_i \in \mathbb{R}, x_i \geq 0$ or $x_i \in \{0,1\}$ for all $i \in [n]$. Creating random real-valued, non-negative, or binary signals of a given sparsity is immediate. To create signals that are sparse in a transformed domain we rely on the work of [Nam13]. Instead of choosing the support of $\overline{x}$ uniformly at random, we choose a set $\Lambda \subseteq [p]$ in order to create a signal with $\Omega_{\Lambda,\bullet}\overline{x} = 0$. Given a subset Λ and a randomly generated vector v with nonzero entries, we obtain the desired signal by calculating

$$\overline{x} = (I - \Omega_{\Lambda,\bullet}^{\top}(\Omega_{\Lambda,\bullet}\Omega_{\Lambda,\bullet}^{\top})^{-1}\Omega_{\Lambda,\bullet})v. \tag{3.81}$$

Since $\Omega\overline{x}$ represents differences of signal entries for consecutive indices, taking the absolute value of each signal entry does not change the cosupport of the resulting signal. Therefore, we use this method to generate non-negative signals. Binary signals are obtained by using Algorithm 3.1.

ℓ_1-Minimization In this section, we compare the existing tight upper bounds for the statistical dimension $\delta(\mathcal{D}(f,\overline{x}))$ for f defined in (3.2a) - (3.2c) to the empirical phase transition obtained by the testset displayed in Figure 3.3.

To this end, we first set the ambient dimension $n = 100$ and generate 10 instances of a sparse one-dimensional signal, see Figure 3.2. The rows of Figure 3.2 display signals $\overline{x}$ with equal number of nonzeros in $\overline{x}$, i.e., $\|\overline{x}\|_0$. The columns of Figure 3.2 display signals that have entries sampled either from $\mathbb{R}^n, \mathbb{R}^n_{\geq 0}$ or $\{0,1\}^n$ from left to right. For each pair of relative sparsity and given number m of random linear measurements we verify whether the signal is the unique solution of (3.2), where f corresponds to one instance of (3.2a) - (3.2c). Uniqueness is tested as described in Section 3.1.3. Using Mosek[4], we solve the occurring optimization problems and plotted a phase transition in Figure 3.6. The gray value shows the empirical probability of uniqueness for each pair of parameters (relative sparsity or relative gradient sparsity, number of measurements):

[4] https://www.mosek.com

Algorithm 3.1: Algorithm for converting a random gradient sparse signal (w.r.t. the one-dimensional discrete gradient) into a binary gradient sparse signal.

Input: $\Lambda \subseteq [n-1]$ desired cosupport
Output: $\overline{x} \in \{0,1\}^n$ with cosupport w.r.t. Λ, i.e $\Omega_\Lambda \overline{x} = 0$

$\overline{x}_0 = 1$
for $i \in [n-1]$ **do**
  if $i \in \Lambda$ **then**
    $\overline{x}_{i+1} \leftarrow \overline{x}_i$
  else
    $\overline{x}_{i+1} \leftarrow 1 - \overline{x}_i$
  end
end

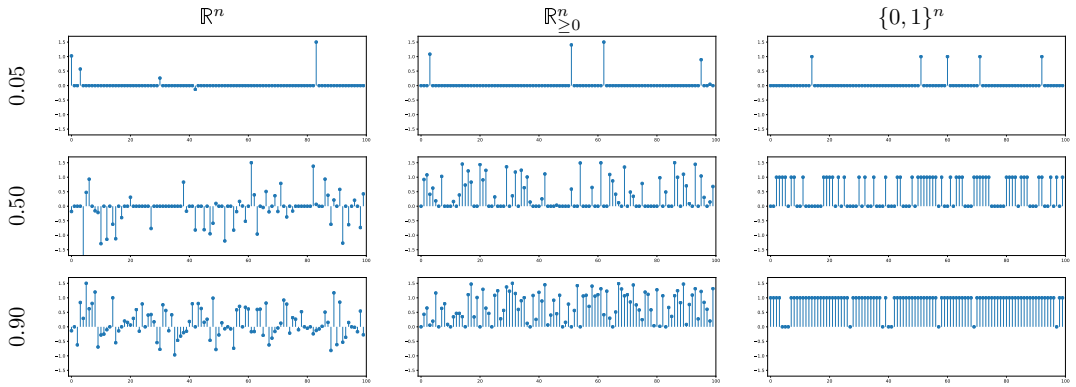


Figure 3.2: Testset for one-dimensional sparse signals. The rows display signals $\overline{x}$ with equal number of nonzeros in $\overline{x}$, i.e., $\|\overline{x}\|_0$. The columns display signals that have entries sampled either from $\mathbb{R}^n, \mathbb{R}^n_{\geq 0}$ or $\{0,1\}^n$ from left to right.

black 0% uniqueness rate, white 100% uniqueness rate. Both regions are accurately separated by our approximation to the statistical dimension. Rows from top to bottom show results for: ℓ_1-, one-dimensional and two-dimensional TV-minimization. Columns from left to right show the signal/image entries: $\mathbb{R}^n, \mathbb{R}^n_{\geq 0}$ or $\{0,1\}^n$. The red curves show $\min_{\tau \geq 0} J(\tau)$ separating these regions accurately.

TV-Minimization To generate the phase diagrams, we proceed as above but use the gradient sparse one-dimensional signals, see Figure 3.3. The rows display signals $\overline{x}$ with equal number of nonzeros in Ωx, i.e., $\|\Omega \overline{x}\|_0$, with respect to the one-dimensional finite difference operator. The columns display signals that have entries sampled either from $\mathbb{R}^n, \mathbb{R}^n_{\geq 0}$ or $\{0,1\}^n$ from left to right. Since explicit curves for J are missing, we use an approximation to the upper bound (3.80). To approximate the expected value,

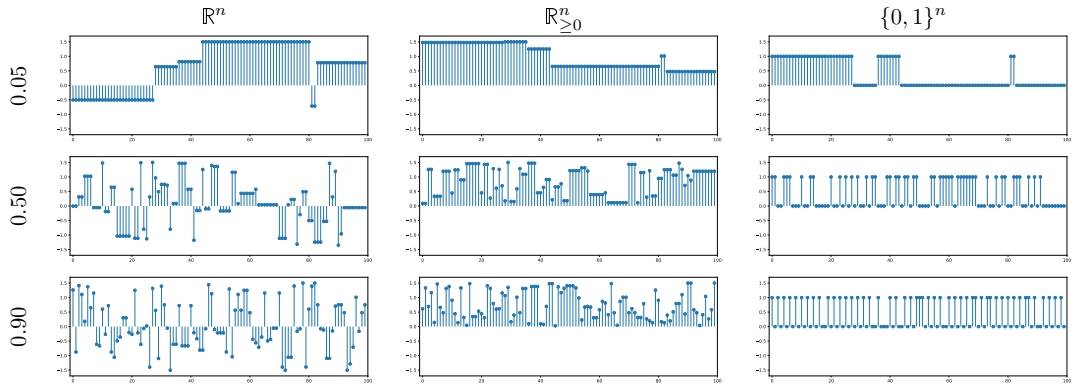


Figure 3.3: Testset for one-dimensional gradient sparse signals. The rows display signals $\overline{x}$ with equal number of nonzeros in $\overline{x}$, i.e., $\|\overline{x}\|_0$. The columns display signals that have entries sampled either from $\mathbb{R}^n, \mathbb{R}^n_{\geq 0}$ or $\{0,1\}^n$ from left to right.

we use a large $k = 10000$ and calculate

$$h_k(\tau) = \frac{1}{k} \sum_{i=1}^{k} \operatorname{dist}(X_i, \tau \partial f(\overline{x}))^2 \approx J(\tau), \tag{3.82}$$

where each X_i is sampled component-wise from the normal distribution $\mathcal{N}(0,1)$. Additionally, we showed in Proposition 3.27 that the unique minimum lies in a compact interval. Therefore, we use the BiSection method [Bar08, Sec. 2.2] to find the minimum of (3.82).

Note that for computing each $\operatorname{dist}(X_i, \tau \partial f(\overline{x}))^2$ one is faced with solving a quadratic program with linear constraints, since $\partial f(\overline{x})$ may formulated as a system of linear equations

$$g \in \partial f(\overline{x}) \Leftrightarrow \exists \alpha : g = \Omega \alpha,\ \alpha_{\Lambda^c} = \operatorname{sign}(\Omega_{\Lambda^c,\bullet} \overline{x}),\ -\mathbb{1} \leq \alpha_\Lambda \leq \mathbb{1}. \tag{3.83}$$

We refer to (3.59) for the definition of $\partial f(\overline{x})$ when $f(x) = \|\Omega x\|_1$.

As for ℓ_1-minimization, we show in Figure 3.6 the phase transition for the one-dimensional TV-minimization. We emphasize the perfect agreement of the phase transition with the approximated upper bound.

3.3.2 2D Empirical Phase Transitions

Testset Creating gradient sparse images with a given gradient sparsity is more involved than in the one dimensional case. By using random support sets $\Lambda \subset [p]$ with $|\Lambda| \geq n$ and the projecting technique (3.81), we most likely obtain constant two-dimensional images. Thus, we use a different approach to construct random gradient sparse images. To this end, we randomly add binary images with homogeneous areas and use the modulo operation to binarize again the result. We show some results in

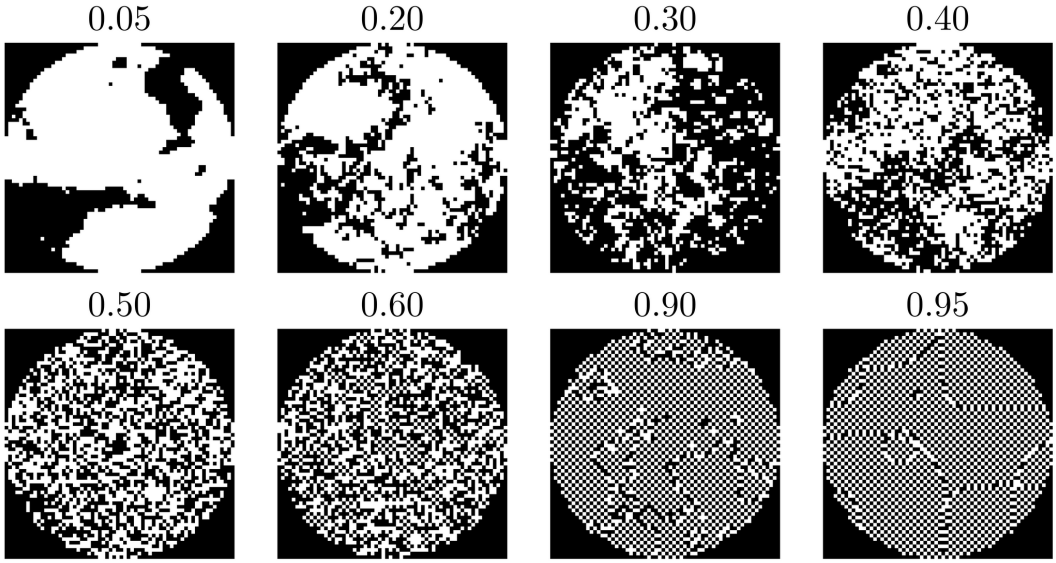


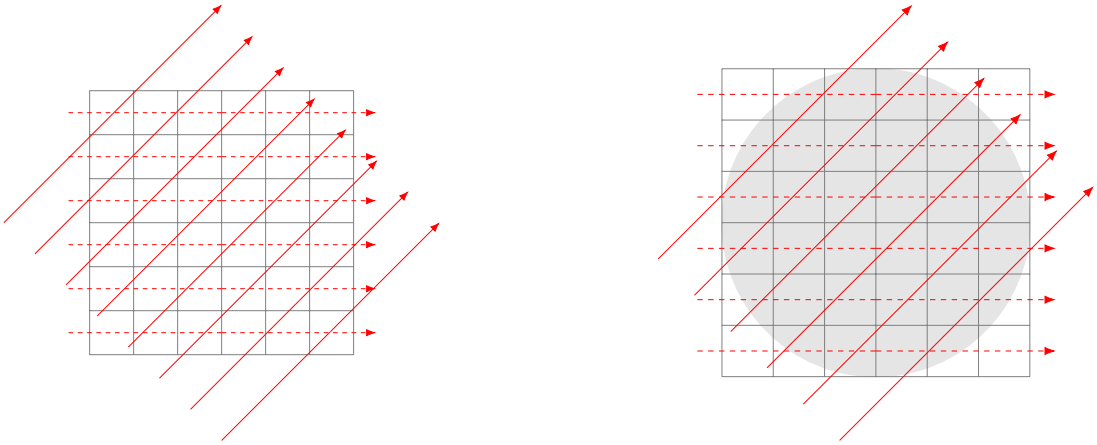
Figure 3.4: Testset for two-dimensional gradient sparse images. Each image label shows the relative sparsity of $\Omega\overline{x}$, i.e., $\|\Omega\overline{x}\|_0/p$, with respect to the two-dimensional finite difference operator. Our test images are 64×64.

Figure 3.4 for two-dimensional gradient sparse images. Each image label shows the relative sparsity of $\Omega\overline{x}$, i.e., $\|\Omega\overline{x}\|_0/p$, with respect to the two-dimensional finite difference operator. Our test images are 64×64.

Creating non-binary gradient sparse images, we identify the connecting components of a binary image. We define a connected component by the induced graph of the binary image as follows. For the definition of a graph, we refer to Section 2.1.3. Let $x \in \{0,1\}^n$ represents a binary image, and $\mathcal{V} = [n]$. The edge set is defined by $ij \in \mathcal{E} \Leftrightarrow x_i = x_j$. Each connected component of the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of pixel indices forming a connected component in the binary image x.

Since we are able to efficiently identify all connected components [KV18, Prop. 2.17] in a binary image, we assign random real values to the different connected components. The resulting discrete gradient of the new image has the same number of nonzeros, but the image is non-binary.

Tomographic Measurements Each projecting direction leads in discrete tomography in general to a different number of measurements. We adopt the idea of [RLH14] that leads to tomographic measurements that carry the same amount of information for each projecting direction/angle. In Figure 3.5, we illustrate the effect on the tomographic projections when sensing an images embedded in a rectangle shape resp. to a circular shape. We emphasize that our testset for 2D contains only images embedded in a circular shape.



(a) rectangle (b) circle

Figure 3.5: Illustration of parallel projections of a rectangle image and of an image embedded in a circular mask. In (a) one needs at different angles a different amount of equidistant parallel rays to cover the whole image. For sensing an image in a circular (b) mask the number of equidistant parallel rays covering the whole image are the same.

TV-Minimization For fixed ambient dimension $n = 64 \cdot 64$, we chose a relative image sparsity $\frac{\|\Omega\overline{x}\|_0}{p} \in [0,1]$, with $\Omega \in \mathbb{R}^{p\times n}$, corresponding to a gradient sparse test image and also choose the number of measurements m. Measurements are Gaussian or tomographic. We consider three types of tomographic matrices:

- binary, as described in [RLH14];
- perturbed binary, that are matrices with the same incidence structure as above but with perturbed nonzero entries in order to remove linear dependencies between rows;
- standard tomographic matrices, with non-negative real entries. We use the MATLAB routine `paralleltomo.m` from the AIR Tools package [HJ18] that implements such a tomographic matrix for an arbitrary vector of angles. We choose equidistant angels, set `N=64` the image size and use the default value of q, i.e., the number of parallel rays for each angle `q = round(sqrt(2)*n)` to obtain a tomographic matrix of size $m \times n$.

Results are presented in Figure 3.7. All plots display a phase transition and thus exhibit regions where exact image reconstruction has empirical probability equal or close to one. All regions are accurately separated by our approximation to the statistical dimension.

3.3.3 Phase Transition Curves in Practice

Figure 3.7 illustrates that the phase transition curves for random measurements obtained from tomographic measurements separate quite accurately success and failure. One use this observation to predict the number of measurements needed to uniquely

reconstruct an image from *tomographic* measurements with high probability. As an example, we depict in Figure 3.8 a way to acquire this information from a phase transition curve. In case of tomography, one has to keep in mind that we usually cannot add row-wise measurements to the matrix since every tomographic projection consists of a fixed number of rows. Taking the example from Figure 3.8 (a), we need at least $256^2 \cdot 0.12$ linear independent rows. Assuming that each tomographic projection contains 256 rows, we need $d \geq 31$ tomographic projections, since

$$\begin{aligned} & 256^2 \cdot 0.12 \leq 256 \cdot d \\ \Leftrightarrow \quad & \frac{256^2 \cdot 0.12}{256} \approx 30.72 \leq d. \end{aligned}$$

We do the same calculation for the image in Figure 3.8 (c) and obtain $d \geq 41$ projections.

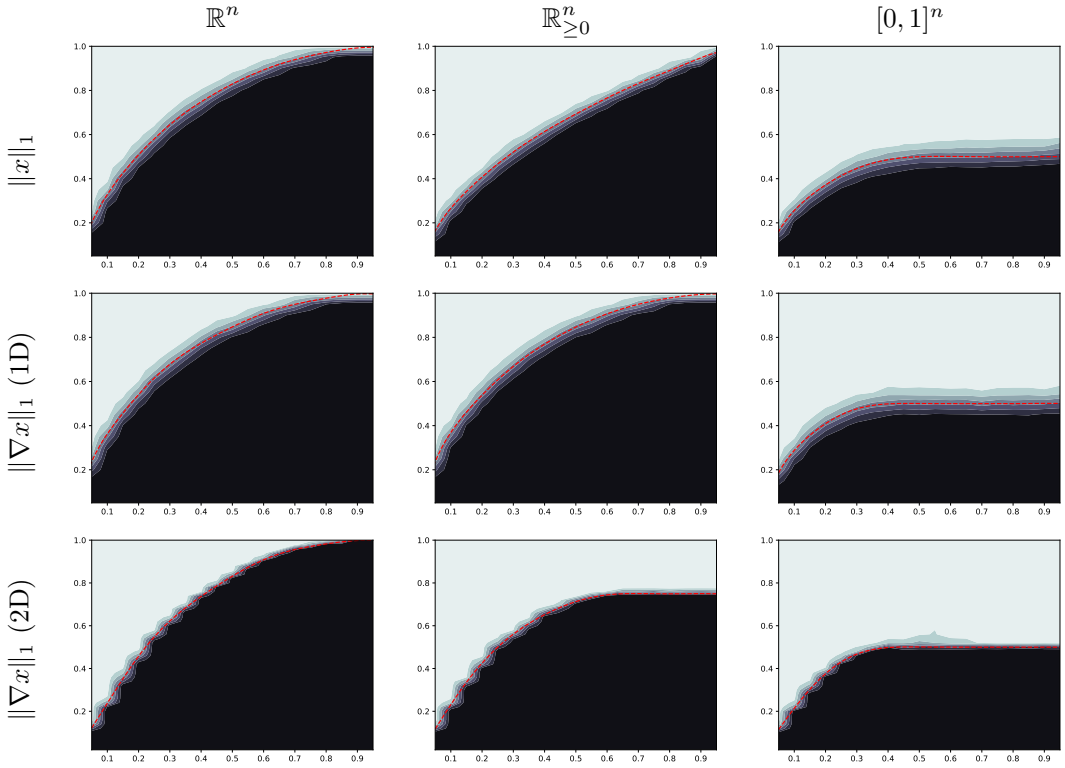


Figure 3.6: The gray value shows the empirical probability of uniqueness for each pair of parameters (relative sparsity or relative gradient sparsity, number of measurements): black 0% uniqueness rate, white 100% uniqueness rate. Both regions are accurately separated by our approximation to the statistical dimension. Rows from top to bottom show results for: ℓ_1-, 1D TV- and 2D TV-minimization. Columns from left to right show the signal/image entries: $\mathbb{R}^n, \mathbb{R}^n_{\geq 0}$ or $\{0,1\}^n$. The red curves show that the bound $\min_{\tau \geq 0} J(\tau)$ from (3.38) separates these regions accurately as predicted in Proposition 3.20.

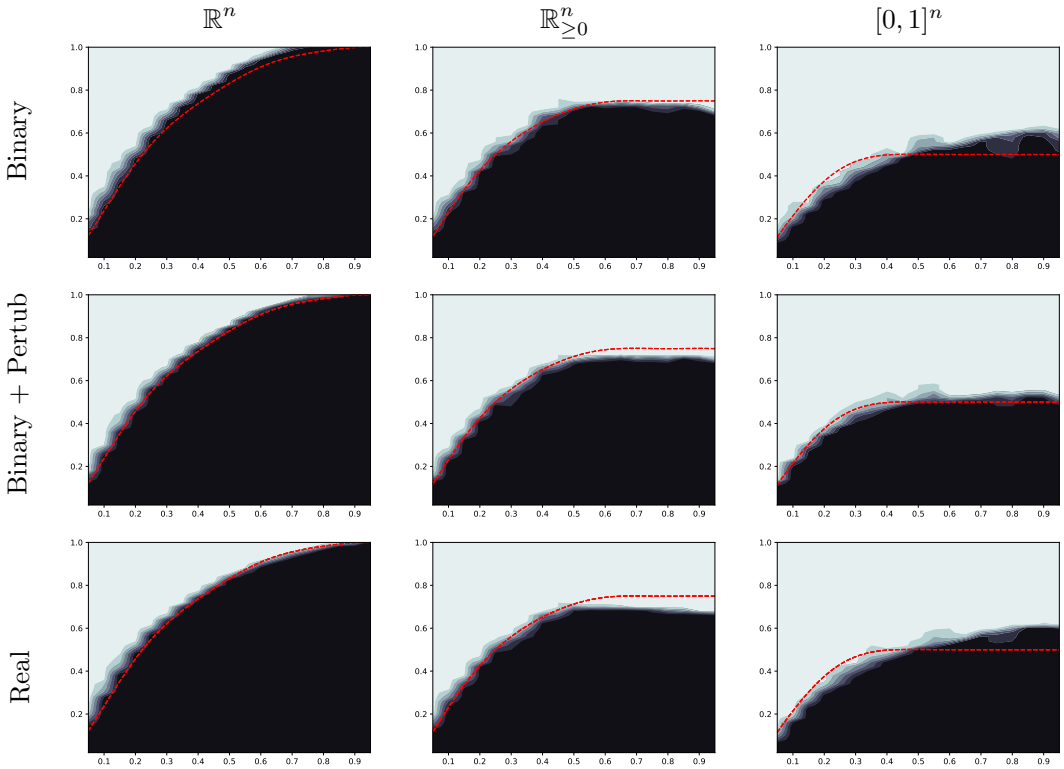
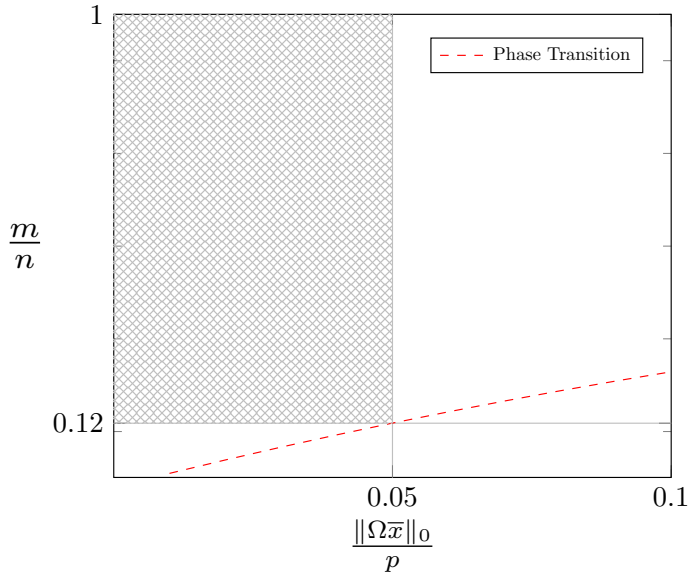


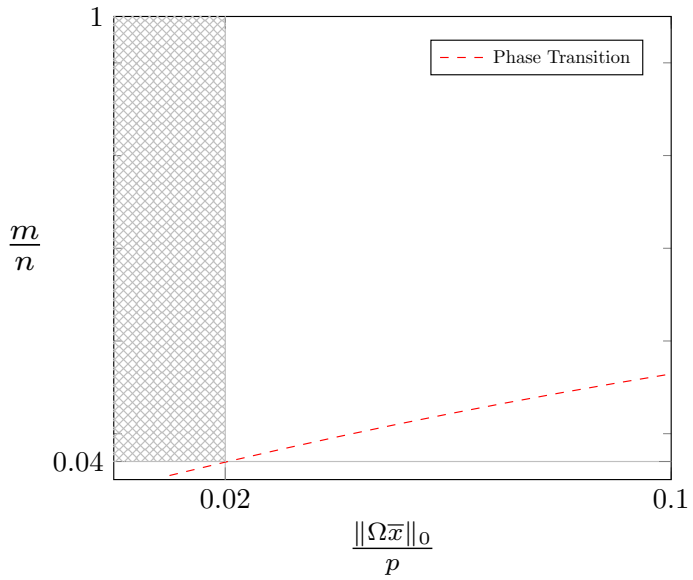
Figure 3.7: We display the empirical probability of uniqueness in the case of *tomographic measurements* for each pair of parameters (relative sparsity or relative gradient sparsity, number of measurements): black 0% uniqueness rate, white 100% uniqueness rate. Both regions are accurately separated by our approximation to the statistical dimension. Rows from top to bottom show results for: binary, perturbed binary and standard matrices, i.e. with non-negative real entries. Columns for left to right show the signal/image entries: $\mathbb{R}^n, \mathbb{R}^n_{\geq 0}$ or $\{0,1\}^n$. The results demonstrate a remarkable agreement of the empirical phase transitions for tomographic recovery with the approximated curve based on the statistical dimension for random measurements.

(a) Binary image with resolution 256×256 and $\frac{\|\Omega\overline{x}\|_0}{p} = 0.05$.



(b) Phase Transition for binary images

(c) Binary image with resolution 1024×1024 and $\frac{\|\Omega\overline{x}\|_0}{p} = 0.02$.



(d) Phase Transition for binary images

Figure 3.8: This figure illustrates the usage of an available phase transition curve. If the relative sparsity of an image is known e.g. (a) and (c), we use the phase transition curve to estimate the minimum number of measurements needed to reconstruct the image uniquely. In (b) and (d) we marked the regions for two different relative sparsities. Conisdering a relative sparsity less or equal than 0.05 like in (a), we see in (b) that every image needs at least $256^2 \cdot 0.12$ linear independend rows in the measuremnt matrix to reconstruct the image uniquely via (3.2f).

CHAPTER 4

A Dual Decomposition Approach

In Chapter 3, we provided lower bounds on the number of measurements required to uniquely recover, by convex programming, an image so that it is gradient sparse and has constrained pixel values. While we considered constrained pixel values in a continuous interval, [Kei17] showed that restricting pixel values to a finite range further reduces the needed number of measurements for unique reconstruction. This motivated us to propose in this Chapter[1] a model that aims at finding

(a) a finite valued image x, .i.e, $x \in \{0, 1, \dots, k-1\}^n$, such that
(b) tomographic projection constraints are given by $Ax = b$ with $A \in \{0,1\}^{m \times n}, b \in \mathbb{N}^m$ are fulfilled, and
(c) x minimizes an *energy* function $E : \{0, 1, \dots, k-1\}^n \to \mathbb{R}$.

For the definition of a binary tomographic matrix A, we refer to Section 2.3.

Regarding computational complexity, even a model considering (a),(b) is NP-hard for more than two tomographic projections [GGP99], implying that a model considering (a)-(c) is too. Thus, it is reasonable to assume that there is no efficient algorithm to find a solution concerning (a)-(c). As a consequence, efficient methods to approximate a solution to (a)-(c) are needed.

We propose in this chapter a novel convex relaxation for (a)-(c) based on linear programming [Van14], which yields a lower bound to the energy to judge the proximity of a solution. Existing convex relaxations [Kap15; Gou13] consider only the binary case, .ie. $x \in \{0,1\}^n$. We prove theoretically and numerically that our relaxation is tighter, hence it better approximates (a)-(c).

To this end, we assume that E factorizes according to a pairwise graphical model [Li09], in our case, a *Markov Random Field*. That is an important class of graph-structured models in image processing, and machine learning: given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, together with a *label space* $\mathcal{X}_\mathcal{V} := \prod_{u \in \mathcal{V}} \mathcal{X}_u$, $\mathcal{X}_u := [k_u]$ for all $u \in \mathcal{V}$, $k_u \in \mathbb{N}$, the

[1] This chapter is based on [KSP17].

energy is a sum of *unary potentials* $\theta_u : \mathcal{X}_u \to \mathbb{R}$ for all $u \in \mathcal{V}$ and *pairwise potentials* $\theta_{uv} : \mathcal{X}_u \times \mathcal{X}_v \to \mathbb{R}$ for all $uv \in \mathcal{E}$. The definition of graphs is provided in Section 2.1.3.

The potentials assign costs to each label $x_u \in \mathcal{X}_u$. In particular, the unary potentials assign costs to label $x_u \in \mathcal{X}_u$ for node $u \in \mathcal{V}$ and the pairwise potentials to a combination $(x_u, x_v) \in \mathcal{X}_u \times \mathcal{X}_v$ of labels for an edge $uv \in \mathcal{E}$. Hence, the entire problem reads

$$\min_{x \in \mathcal{X}_\mathcal{V}} E(x) := \sum_{u \in \mathcal{V}} \theta_u(x_u) + \sum_{uv \in \mathcal{E}} \theta_{uv}(x_u, x_v) \quad \text{s.t.} \quad Ax = b\,. \tag{4.1}$$

For the discrete tomography problem, we usually choose $\mathcal{G}$ to be a grid graph corresponding to the pixels of the image to be reconstructed, zero unary potentials $\theta_u \equiv 0$ for all $u \in \mathcal{V}$, **as no local information about the image values is known**. Pairwise potentials $\theta_{uv} = g(x_u - x_v)$ penalize intensity transitions, e.g., the choice $g(\cdot) = |\cdot|$ leads to TV-regularization and $g(\cdot) = \min(1, |\cdot|)$ to Potts-regularization. Such choice of pairwise potentials assigns small energy to labelings x with a regular spatial structure.

The outline of this chapter is as follows. First, we show that if the linear constraint in (4.1) is a single row corresponding to a single tomographic projection ray, then (4.1) is efficiently solvable. Decomposing (4.1) into subproblems consisting only of a single tomographic projection ray, we use Lagrange multipliers to enforce consistency between solutions of the subproblems. This enables us to construct a graphical model with higher-order terms, but without additional linear constraints. Further, we introduce *message passing* [Kol15] as an efficient algorithm to exactly solve the occurring subproblems. Finally, we prove theoretically and numerically that our proposed model yields a tighter relaxation than existing approaches.

4.1 One-Dimensional Non-Binary Discrete Tomography

Our decomposition of the discrete tomography problem (4.1) consists of considering a subproblem for each ray constraint separately and joining them together via Lagrangian variables.

In particular, let $U = \{u_1, \dots, u_n\} \subseteq \mathcal{V}$ be the variables from a single ray constraint $x_{u_1} + \dots + x_{u_n} = b$ corresponding to a row of the projection matrix A in (4.1). For each pair $u_i, u_{i+1} \in U$ with $i \in [n-1]$, we assume that u_i and u_{i+1} are adjacent, i.e., $u_i u_{i+1} \in \mathcal{E}$. In terms of graph theory, the sub-graph induced by $\mathcal{G}' = (U, (U \times U) \cap \mathcal{E})$ is called a *chain*. Then, we define the *one-dimensional discrete tomography problem* by

$$\begin{aligned} \min_{(x_1,\dots,x_n) \in \mathcal{X}_U} & \sum_{u \in U} \theta_u(x_u) + \sum_{i \in [n-1]} \theta_{u_i, u_{i+1}}(x_{u_i}, x_{u_{i+1}}) \\ \text{s.t.} & \sum_{u \in U} x_u = b, \end{aligned} \tag{4.2}$$

that involves only variables corresponding to a single ray constraint. In the following, we present an exact linear programming relaxation and an efficient message-passing routine to solve (4.2).

4.1.1 Linear Programming Model

The one-dimensional discrete tomography subproblem (4.2) could naively be solved by *dynamic programming* [Bel54] by going over all variables $u_1, \dots, u_n$ sequentially.

Dynamic programming is a well-known approach to problems of *combinatorial optimization* [KV18]. One assumes that the problem is hierarchically decomposable and the overall solution is constructed by the intermediate results of the sub-problems.

Below, we use the *big O notation* [AB16] to compare approaches for dynamic programming. For two functions $f : \mathbb{N} \to \mathbb{N}$ and $g : \mathbb{N} \to \mathbb{N}$ the *big O notation* is defined by

$$f = O(g) \Leftrightarrow \exists c > 0 \; \exists n_0 \; \forall n > n_0 : f(n) \leq c \cdot g(n). \tag{4.3}$$

Writing the number intermediate results increases by $O(g)$ means that for the actual function f describing the increase it holds $f = O(g)$.

Solving (4.2) by dynamic programming, we calculate for each variable u_i the costs for all possible labels x_{u_i} and all values of the intermediate sum $\sum_{j=1}^{i-1} x_{u_j}$. The latter sum has $1 + \sum_{j=1}^{i-1}(|\mathcal{X}| - 1)$ possible values. In total, the number of intermediate results increases quadratic $O(n^2)$ with respect to the number of nodes n to construct a solution for (4.2).

To reduce the number of intermediate results, we recursively

(i) equipartition variables $u_1, \dots, u_n$,

(ii) define LP-subproblems in terms of so-called *counting factors* which are exact on each subpartition, and

(iii) join them to eventually obtain an exact LP-relaxation for (4.2).

Our approach is inspired by [Tar12].

Partition of variables. Given the nodes $u_1, \dots, u_n$, we choose an equipartition $\Pi_1 = \{u_1, \dots, u_{\lfloor n/2 \rfloor}\}$ and $\Pi_2 = \{u_{\lfloor n/2 \rfloor + 1}, \dots, u_n\}$. We recursively equipartition Π_1 into $\Pi_{1,1}$ and $\Pi_{1,2}$ and do likewise for Π_2. For $u_1, \dots, u_8$, we obtain a recursive partitioning as in Figure 4.1.

Counting factors. Given an interval, a *counting factor* holds the states of its left and right end and the value of the intermediate sum.

Definition 4.1 (Counting label space)**.** The counting label space for interval $[i : j]$ is $\mathcal{X}_{i:j} := \mathcal{X}_{u_i} \times \mathcal{S}_{i:j} \times \mathcal{X}_{u_j}$ with $\mathcal{S}_{i:j} = [1 + \sum_{l=i+1}^{j-1}(|\mathcal{X}_{u_l}| - 1)]$ holding all possible intermediate sums. A *counting label* $x_{i:j}$ consists of the three components $(x_{u_i}, s_{i:j}, x_{u_j})$: its left endpoint label x_{u_i}, intermediate sum $s_{i:j} := x_{u_{i+1}} + \ldots + x_{u_{j-1}}$ and right endpoint label x_{u_j}.

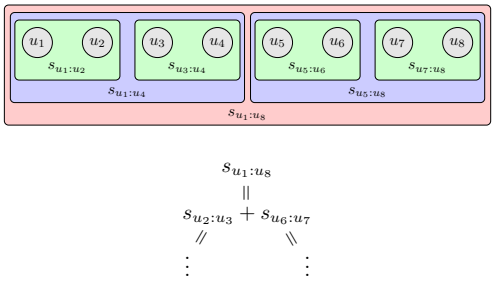


Figure 4.1: *Partition of variables*: The chain of variables $U = \{u_1, \dots, u_8\}$ (red box) is first partitioned into $\Pi_1 = \{u_1, u_2, u_3, u_4\}, \Pi_2 = \{u_5, u_6, u_7, u_8\}$ (blue boxes), and then into $\Pi_{1,1} = \{u_1, u_2\}, \Pi_{1,2} = \{u_3, u_4\}, \Pi_{2,1} = \{u_5, u_6\}, \Pi_{2,2} = \{u_7, u_8\}$ (green boxes).
Counting sums: The topmost sum $s_{u_1:u_8} = x_{u_2} + \ldots + x_{u_7}$ is composed of sums $s_{1:4}$ for partition Π_1 and $s_{5:8}$ for Π_2. Subsums are generated recursively for $s_{1:4}$ and $_{5:8}$ again.

See Figure 4.1 for the exemplary case $U = \{u_1, \dots, u_8\}$.

For interval $[i:j]$ there are $|\mathcal{X}_{i:j}| = |\mathcal{X}_{u_i}| \cdot |\mathcal{X}_{u_j}| \cdot |\mathcal{S}_{i:j}|$ distinct counting labels. We associate to each counting factor *counting marginals* $\mu_{i:j}$ satisfying $\{\mu_{i:j} \in \mathbb{R}_+^{|\mathcal{X}_{i:j}|} : \sum_{x_{i:j} \in \mathcal{X}_{i:j}} \mu_{i:j}(x_{i:j}) = 1\}$.

Assuming uniform label spaces $|\mathcal{X}_u| = k$ for all $u \in \mathcal{V}$, the total size of all counting factors is $O(k^3 \cdot n \cdot \log(n))$, hence subquadratic in the number of nodes in U.

Joining counting factors. Assume the partitioning of variables has produced two adjacent subsets $\Pi = \{u_i, \dots, u_j\}$ and $\Pi' = \{u_{j+1}, \dots, u_l\}$, which were constructed from their common subset $\Pi \cup \Pi' \subseteq U$. The associated three counting factors with marginals $\mu_{i:j}, \mu_{j+1:l}$, and $\mu_{i:l}$ introduced above should be consistent with each other.

Definition 4.2 (Label consistency)**.** Label $x_{i:l} \in \mathcal{X}_{i:l}$, $x_{i:j} \in \mathcal{X}_{i:j}$ and $x_{j+1:l} \in \mathcal{X}_{j+1:l}$ are *consistent* with each other, denoted by $x_{i:j}, x_{j+1:l} \sim x_{i:l}$ if and only if

(i) left endpoint labels of $x_{i:j}$ and $x_{i:l}$ are equal,
(ii) right endpoint labels of $x_{j+1:l}$ and $x_{i:l}$ are equal and
(iii) intermediate sums are equal $s_{i:l} = s_{i:j} + x_{u_j} + x_{u_{j+1}} + s_{j+1:l}$.

We enforce this by introducing a *higher-order marginal* $\mu_{i:j:l} \in \mathbb{R}_+^{\mathcal{X}_{i:j} \times \mathcal{X}_{j+1:l}}$ to bind together $\mu_{i:j}, \mu_{j+1:l}$ and $\mu_{i:l}$.

$$\sum_{x_{j+1:l}} \mu_{i:j:l}(x_{i:j}, x_{j+1:l}) = \mu_{i:j}(x_{i:j}), \qquad \forall x_{i:j} \in \mathcal{X}_{i:j}, \tag{4.4}$$

$$\sum_{x_{i:j}} \mu_{i:j:l}(x_{i:j}, x_{j+1:l}) = \mu_{j+1:l}(x_{j+1:l}), \qquad \forall x_{j+1:l} \in \mathcal{X}_{j+1:l}, \tag{4.5}$$

$$\sum_{x_{i:j}, x_{j+1:l} \sim x_{i:l}} \mu_{i:j:l}(x_{i:j}, x_{j+1:l}) = \mu_{i:l}(x_{i:l}), \qquad \forall x_{i:l} \in \mathcal{X}_{i:l}\,. \tag{4.6}$$

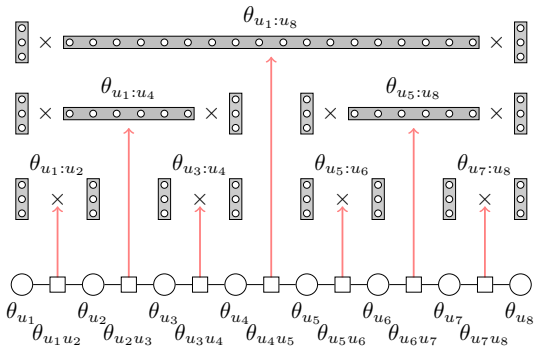


Figure 4.2: Illustration of the construction used for solving the one-dimensional tomography problem (4.2) with $\ell = 3$, i.e., three labels. On the bottom is the original chain; above, there are three layers of counting factors. Each counting factor keeps track of the label x_u (left vertical bar), x_v (right vertical bar), and the intermediate sum $\sum_{w \in U: u<w<v} x_w$ (middle horizontal bar).

The recursive arrangement of counting factors is illustrated in Figure 4.2.

Remark 4.1. The constraints between $\mu_{i:j:l}$ and $\mu_{i:j}$ and $\mu_{j+1:l}$ are analogous to the marginalization constraints between pairwise and unary marginals in the local polytope relaxation for pairwise graphical models [Wer07]. The constraints between $\mu_{i:j:l}$ and $\mu_{i:l}$, however, are different. Hence, specialized, efficient solvers for inference in graphical models cannot be applied.

Costs. Above, we have described the polytope for the one-dimensional discrete tomography problem (4.2). The LP-objective consists of vectors $\theta_{i:j}$ for each counting marginal and $\theta_{i:j:l}$ for each higher-order marginal. Accounting for the pairwise costs in (4.2) we define

$$\theta_{i:j}(x_{i:j}) := \begin{cases} \theta_{u_i u_j}(x_{u_i}, x_{u_j}), & i+1 = j \\ 0, & \text{otherwise} \end{cases} \tag{4.7}$$

for the counting factors and for the higher-order factors we define

$$\theta_{i:j:l}(x_{i:j}, x_{j+1:l}) := \theta_{u_j u_{j+1}}(x_{u_j}, x_{u_{j+1}}). \tag{4.8}$$

For the projection constraint in (4.2) we set costs of the top counting marginal by

$$\theta_{1:n}(x_{1:n}) := \begin{cases} 0, & x_{u_1} + s_{1:n} + x_{u_n} = b \\ \infty, & \text{otherwise} \end{cases}. \tag{4.9}$$

4.1.2 Message Passing Algorithm

In Section 4.1.1, we have introduced a linear program formulation for the one-dimensional discrete tomography problem (4.2). While it is possible to solve it with a standard LP-solver, doing so would be slow. As the counting factors and the higher-order marginals connecting them form a graph containing no circles, called a *tree* in graph theory, it is possible to devise a message-passing algorithm that solves (4.2) exactly. First, this implies that the linear programming relaxation for (4.2) is exact, as message passing amounts to optimizing the Lagrangian dual of this same relaxation. Second, marginals do not need to be held explicitly; holding messages is enough. The size of all messages equals the size of all counting factors.

Message passing for (4.2) is detailed in Algorithm 4.1. It proceeds by first computing *up messages* from adjacent fine subsets to coarser subsets (i.e., going up the tree in Figure 4.2) and afterward computing *down messages* from coarse subsets to their equipartition (i.e., going down the tree in Figure 4.2). Messages reparametrize the costs of counting and higher-order factors.

Reparametrization Let indices $i < j < l$ be given, where $[i:j]$, $[j+1:l]$ and $[i:l]$ are subsets generated by the recursive partitioning. Then, we define reparametrizations of the potentials $\theta_{i:j:l}(x_{i:j}, x_{j+1:l}), \theta_{i:j}(x_{i:j}), \theta_{j+1:l}(x_{j+1:l})$ and $\theta_{i:l}(x_{i:l})$ by

$$\theta^{\phi}_{i:j:l}(x_{i:j}, x_{j+1:l}) := \theta_{i:j:l}(x_{i:j}, x_{j+1:l}) + \phi_{i:j:l}(x_{i:j}) + \phi_{i:j:l}(x_{j+1:l}) + \phi_{i:j:l}(x_{i:l}), \quad (4.10)$$

$$\theta^{\phi}_{i:j}(x_{i:j}) := \theta_{i:j}(x_{i:j}) - \phi_{i:j:l}(x_{i:j}), \quad (4.11)$$

$$\theta^{\phi}_{j+1:l}(x_{j+1:l}) := \theta_{j+1:l}(x_{j+1:l}) - \phi_{i:j:l}(x_{j+1:l}), \quad (4.12)$$

$$\theta^{\phi}_{i:l}(x_{i:l}) := \theta_{i:l}(x_{i:l}) - \phi_{i:j:l}(x_{i:l}), \quad (4.13)$$

where the values of ϕ correspond to the Lagrange multipliers of the constraints (4.4), (4.5) and (4.6) respectively. Messages $\phi^{\leftarrow}_{i:j:l}, \phi^{\rightarrow}_{i:j:l}, \phi^{\uparrow}_{i:j:l}$ act on (reparametrize) costs θ^{ϕ} by updating the corresponding Lagrange multipliers.

„Sending a message“ in Algorithm 4.1, e.g., for $\phi^{\leftarrow}_{i:j:l}$, implies the creation of new multipliers by.

$$\phi'_{i:j:l}(x_{i:j}, x_{j+1:l}) = \phi_{i:j:l}(x_{i:j}) + \phi^{\rightarrow}_{i:j:l}(x_{i:j}), \quad (4.14)$$

$$\phi'_{i:j}(x_{i:j}) = \phi_{i:j}(x_{i:j}) - \phi^{\rightarrow}_{i:j:l}(x_{i:j}). \quad (4.15)$$

For the description of „sending a message“, we introduce the notation „+=“ corresponding to (4.14) and „−=“ corresponding to (4.15). Additionally, we take a computational point of view by reusing ϕ instead of writing ϕ'. Hence, we provide all message

updates used in Algorithm 4.1 by

$$\theta^{\phi}_{i:j:l}(x_{i:j}, x_{j+1:l}) \mathrel{+}= \phi^{\leftarrow}_{i:j:l}(x_{i:j}), \qquad \theta^{\phi}_{i:j}(x_{i:j}) \mathrel{-}= \phi^{\leftarrow}_{i:j:l}(x_{i:j}), \tag{4.16}$$

$$\theta^{\phi}_{i:j:l}(x_{i:j}, x_{j+1:l}) \mathrel{+}= \phi^{\rightarrow}_{i:j:l}(x_{j+1:l}), \qquad \theta^{\phi}_{j+1:l}(x_{j+1:l}) \mathrel{-}= \phi^{\rightarrow}_{i:j:l}(x_{j+1:l}), \tag{4.17}$$

$$\theta^{\phi}_{i:j:l}(x_{i:j}, x_{j+1:l}) \mathrel{+}= \phi^{\uparrow}_{i:l}(x_{i:l}), \qquad \theta^{\phi}_{i:l}(x_{i:l}) \mathrel{-}= \phi^{\uparrow}_{i:j:l}(x_{i:l}). \tag{4.18}$$

Algorithm 4.1: Message passing for one-dimensional discrete tomography

1 **Up messages:**
2 **for** $[i:j] \cup [j+1:l] = [i:l] \in \Pi$ *in ascending order*
3 $\quad \phi^{\leftarrow}_{i:j:l} = \theta_{i:j}$
$\quad \phi^{\rightarrow}_{i:j:l} = \theta_{j+1:l}$ `// Send messages to higher-order counting factor` $\theta^{\phi}_{i:j:l}$
4 $\quad \phi^{\uparrow}_{i:j:l}(x_{i:l}) = \min_{x_{i:j}, x_{j+1:l} \sim x_{i:l}} \theta^{\phi}_{i:j:l}(x_{i:j}, x_{j+1:l})$ `// Send message from higher order to counting factor` $\theta^{\phi}_{i:l}$
5 **end**
6 $x^*_{1:n} \in \arg\min_{x_{1:n} \in \mathcal{X}_{1:n}} \theta_{1:n}(x_{1:n})$; `// optimum of top counting factor`
7 **Down messages:**
8 **for** $[i:j] \cup [j+1:l] = [i:l] \in \Pi$ *in descending order*
9 $\quad \phi^{\uparrow}_{i:j:l} = \theta_{i:j:l}$; `// Send message to higher-order factor`
10 $\quad x^*_{i:j}, x^*_{j+1:l} \in \underset{x_{i:j}, x_{j+1:l} \sim x^*_{i:l}}{\arg\min} \theta^{\phi}_{i:j:l}(x_{i:j}, x_{j+1:l})$; `// compute optimal labels`
11 **end**

Fast message computation Naively computing one „up messages“ would result in time complexity $O(\ell^5 \cdot n^2)$, which would make the algorithm unacceptably slow. We describe a fast message computation technique for line 3 in Algorithm 4.1, which uses the structure of the corresponding linear constraints (4.6) and relies on the latent factorization of $\theta^{\phi}_{i:j:l}$. Specifically, when we fix the endpoints x_{u_i}, x_{u_j} of interval $[i:j]$ and $x_{u_{j+1}}, x_{u_l}$ of $[j+1:l]$, (3) becomes

$$\phi^{\uparrow}_{i:j:l}(x_{i:l}) = \min_{s_{i:j} + s_{j+1:l} = s_{i:l} - x_{u_j} - x_{u_j}} \theta_{u_j, u_{j+1}}(x_{u_j}, x_{u_{j+1}}) + \phi^{\leftarrow}_{i:j}(x_{i:j}) + \phi^{\rightarrow}_{j+1:l}(x_{j+1:l}) \,, \tag{4.19}$$

Problem (4.19) is an instance of the min-sum convolution problem: Given $a, b \in \mathbb{R}^n$, compute $c \in \mathbb{R}^{2n-1}$, where $c_i = \min_{j \le i}(a_j + b_{i-j})$. This is by replacing $\phi^{\leftarrow}$ by a, $\phi^{\rightarrow}$ by b, and noting that $\theta_{u_j, u_{j+1}}$ is a constant, as x_{u_j} and $x_{u_{j+1}}$ were fixed. For the min-sum convolution problem efficient algorithms [Bus94] were proposed with expected running time $O(n \log(n))$ under the assumption that sorting a and b results in permutations occurring with uniform probability. Problem (4.19) is efficiently be computed by performing $O(\ell^4)$ min-sum convolutions (one convolution for every choice of endpoints).

Remark 4.2 (Comparison to [Tar12])**.** While our approach for solving (4.2) is inspired by [Tar12], it is notably different: (a) our model includes pairwise potentials forming a chain, while [Tar12] assumes that pairwise potentials do not occur between neighboring subsets. This necessitates storing left and right endpoints in counting factors. (b) [Tar12] optimizes a different objective: they solve the sum-product version of (4.2) (i.e., they exchange „min“by „+“and „+“by „·“in (4.2)). This allows [Tar12] to use fast Fourier transforms for message computations, instead of the harder min-sum convolution problems.

4.2 Graphical Model for Discrete Tomography

The discrete tomography problem (4.1) consists of m distinct one-dimensional subproblems (4.2), where m is the number of rows of the tomographic matrix A. We connect all subproblems (4.2) via Lagrangian variables into one large problem. This procedure is called dual decomposition, see [SGJ11] for an introduction. Specifically, in our discrete tomography problems, subproblems only share variables $v \in \mathcal{V}$, but not edges $e \in \mathcal{E}$ (shared edges are handled analogously). Then for each node $u \in \mathcal{V}$, which participates in the i-th subproblem, we introduce the Lagrangian variable $\lambda_{i,u} \in \mathbb{R}^{|\mathcal{X}_u|}$. The i-th subproblem then consists of solving (4.2) with the subset of variables U_i, where the unary potentials are the Lagrangian variables $\theta_u = \lambda_{i,u}$. We denote its energy by $E_i(\cdot, \lambda_i)$. The overall problem is

$$\max_{\lambda_1,\dots,\lambda_m} \sum_{i=1}^{m} \min_{x \in \mathcal{X}_{U_i}} E_i(x, \lambda_i) \quad \text{s.t.} \quad \sum_i \lambda_{i,u} \equiv 0 \quad \forall u \in \mathcal{V} \,. \tag{CTG}$$

An exemplary 4×4 model with eight subproblems coming from two projection directions is depicted in Figure 4.3.

Optimization of relaxation (CTG)**.** We use *bundle methods* [HL93b] minimizing a non-differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$ by aggregating several subgradients to obtain a descent direction in each iteration. This is an advantage compared to *subgradient methods* [HL93b] since subgradients do not guarantee descent along their direction. In particular, to solve (CTG), we use the bundle solver ConicBundle[2] to find optimal Lagrange multipliers λ and Algorithm 4.1 to find solutions to the one-dimensional subproblems. Since (CTG) is the dual problem of the relaxation of (4.1), the solution from (CTG) yields a lower bound to the optimal value of (4.1).

Primal solution. To obtain a feasible reconstruction, we solve a reduced problem by excluding labels with a high cost: Given dual variables λ_i, let x^* be the optimal solution to the i-th subproblem on variables $U_i \subseteq \mathcal{V}$. For each label $x_u \in \mathcal{X}_u, u \in U_i, x_u \neq x_u^*$ we compute the energy $x'^* \in \arg\min_{\{x' \in \mathcal{X}_U : x_u = x'_u\}} E_i(x', \lambda_i)$ of the minimal reconstruction for subproblem i when the label at u is fixed to x_u (this value can be read off from the

[2] https://www-user.tu-chemnitz.de/~helmberg/ConicBundle/

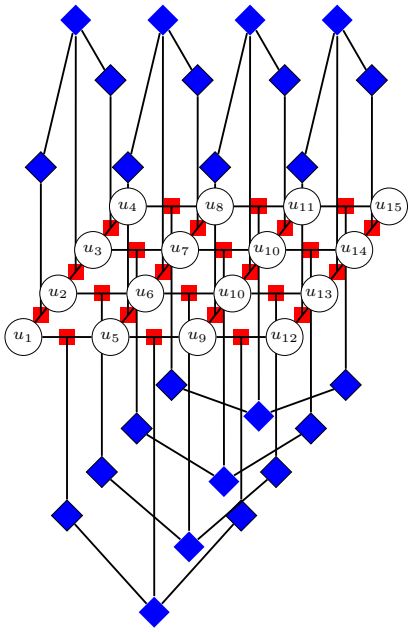


Figure 4.3: This figure is an illustration of a complete graphical model for the discrete tomography problem with projections along rows and columns of the underlying 4×4 grid. The circles (unary potentials) and red squares (pairwise potentials) display the energy function from (4.1) according to the underlying graph. The blue diamonds indicate the higher-order potentials, which we defined to incorporate the linear constraint $Ax = b$ from (4.1) into a graphical model. We depicted two tomographic projections, four vertical and four horizontal rays.

reparametrization output by Algorithm 4.1). Only if the gap $E_i(x'^*, \lambda_i) - E_i(x^*, \lambda_i)$ is smaller than some given threshold, we consider the label x_u. We construct the discrete tomography problem on this reduced set of possible labelings and solve the problem with CPLEX[3].

Comparison to previously used relaxation. The algorithms in [Kap15; Zis16; WSH03; Bat08] use the following relaxation.

$$\begin{aligned} \min_{\mu \geq 0} \quad & \textstyle\sum_{u \in \mathcal{V}} \langle \theta_u, \mu_u \rangle + \sum_{uv \in \mathcal{E}} \langle \theta_{uv}, \mu_{uv} \rangle \\ \text{s.t.} \quad & \textstyle\sum_{x_u \in \mathcal{X}_u} \mu_u(x_u) = 1 \quad \forall u \in \mathcal{V} \\ & \textstyle\sum_{x_u \in \mathcal{X}_u} \mu_{uv}(x_u, x_v) = \mu_v(x_v) \quad \forall x_v \in \mathcal{X}_v \qquad \forall uv \in \mathcal{E} \\ & \textstyle\sum_{x_v \in \mathcal{X}_v} \mu_{uv}(x_u, x_v) = \mu_u(x_u) \quad \forall x_u \in \mathcal{X}_u \\ & \textstyle\sum_{u \in \mathcal{V}} A_{iu} \cdot \left(\sum_{x_u \in \mathcal{X}_u} x_u \cdot \mu_u(x_u) \right) = b_i \quad i = 1, \dots, m \,. \end{aligned} \tag{STD}$$

[3] https://www.ibm.com/de-de/analytics/cplex-optimizer

This relaxation (STD) is the straightforward generalization of the local polytope relaxation [Wer07] to the discrete tomography problem. The only differences are the linear constraints in the last line of (STD). Concerning the one-dimensional discrete tomography problem (4.2), the difference between (STD) and our approach is: for (STD) the tomographic projections are directly enforced through the unary marginals $\mu_u, u \in \mathcal{V}$ instead of enforcing them through the counting factors and higher-order ones as we did in Section 4.1. This more simplistic relaxation (STD) is, however, less tight.

Proposition 4.3. *Relaxation* (STD) *is less tight than* (CTG).

Proof. Relaxation (STD) is equivalent to applying it to each tomographic projection separately and then joining every subproblem by Lagrangian variables as we did with our approach above (CTG), see [SGJ11, Section 1.6]. Hence, it is sufficient to show that (STD) is not tight in the one-dimensional case (4.2). We give a counter-example. Assume $\mathcal{X}_u = \{0, 1\}\ \forall u \in U$ and we are given Potts pairwise potentials

$$\theta_{uv}(x_u, x_v) = \begin{cases} 0, & x_u = x_v \\ 1, & x_u \neq x_v \end{cases} \tag{4.20}$$

and zero unary potentials $\theta_u \equiv 0$. Set unary marginals $\mu_u(1) = \frac{b}{|U|}$ and $\mu_u(0) = 1 - \mu_u(1)\ \forall u \in U$ and pairwise marginals as

$$\mu_{uv}(x_u, x_v) = \begin{cases} \mu_u(x_u), & x_u = x_v \\ 0, & x_u \neq x_v \end{cases}. \tag{4.21}$$

Such marginals are feasible to (STD), yet give cost 0. On the other hand for e.g. $b = 1$ and $|U| > 1$ there must be at least one label transition, which the Potts potential penalizes with cost 1. □

4.3 Experiments

Test images. We used 200 randomly generated 32×32 images with three distinct intensity values $\{0, 1, 2\}$, examples of which are displayed in Figure 4.4. The binary matrix $A \in \{0, 1\}^{m \times n}$ for the tomographic projections were constructed as in Section 2.3. For each test image, we consider two tomographic problems: (i) measuring along horizontal and vertical directions or (ii) measuring along horizontal, vertical, and two diagonal directions (left upper to the right lower and left lower to right upper corner). This gives 400 test problems in total. Potentials for energy E in (4.1) are: unary potentials are zero, while pairwise ones are $\theta_{uv} = |x_u - x_v|$ (that corresponds to TV). Due to the integrality of all costs, optimality is ascertained through a duality gap < 1.

Algorithms. We identify our solvers by a prefix **{CTG|STD}** depending on whether (CTG) or (STD) is solved and by a suffix **{CB|relax|BB}** depending on whether Con-

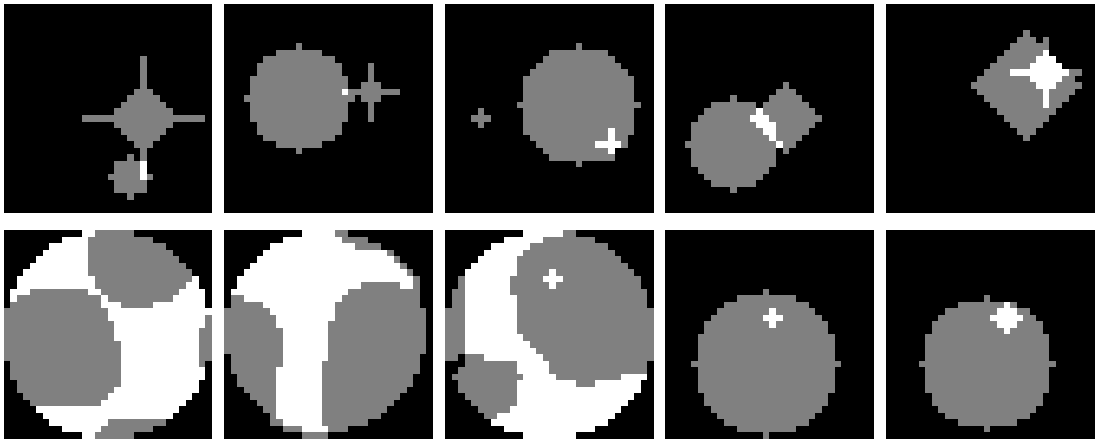


Figure 4.4: These images are examples from our testset of 200 images with size 32×32. Black represents value 0, gray 1 and white 2.

icBundle, CPLEX[4] or CPLEX with branch and bound enabled was utilized. This gives five solvers in total: **CTG_CB**, **CTG_relax**, **CTG_BB**, **STD_relax**, and **STD_BB**. We set a time limit of one hour for all algorithms.

Unfortunately, CPLEX cannot solve problems larger than 32×32 in our scenario. When solving the relaxation (CTG), it already consumes multiple GB of memory for 32×32 images. Solving (STD), on the other hand, leads to low memory consumption, but CPLEX takes too much time for larger problems (> 1 hour). Consequently, we decided to stick with 32×32 images to obtain a baseline.

Results. In Proposition 4.3, we proved that relaxation (STD) is less tight than our relaxation (CTG). Indeed, the first line in Table 4.2 shows that this occurs 350 times. Furthermore, our tighter relaxation helps in giving optimality certificates. In Table 4.1, we confirm this numerically: **STD_relax** yields optimality certificates 53 times, while **CBC_CB** and **CTG_relax** do so 205 times in total.

Interestingly, when using the branch and bound capabilities of CPLEX, the picture changes, and **STD_BB** outperforms **CTG_BB**. This is probably due to the fact that CPLEX solves the underlying relaxation (STD) much faster than (CTG). We conjecture that the picture changes if a more efficient implementation **CBC_CB** is used as a bounds provider inside a branch and bound solver. However, this is beyond the scope of our work.

In Figure 4.5, we give a detailed plot of how much our relaxation (CTG) improved compared to (STD).

Also, our relaxation helps in reconstructing the signal. Out of 238 instances, where our heuristic finds an optimal integral solution (third line in Table 4.2), there are 12 cases, where only our heuristic do so (second line in Table 4.2).

[4] https://www.ibm.com/de-de/analytics/cplex-optimizer

	STD relax	STD BB	CTG CB	CTG relax	CTG BB
duality gap "< 1"	53	243	178	154	182
			205		

Table 4.1: Number of instances where duality gap < 1 (optimality).

	#Instances
(CTG) $>$ (STD) (our relaxation yields **strictly better** lower bound)	350
our heuristic **(only)** found **optimal** integral solution	12
our heuristic found **optimal** integral solution	238

Table 4.2: Comparison of bounds and primal solutions obtained by (STD) or (CTG).

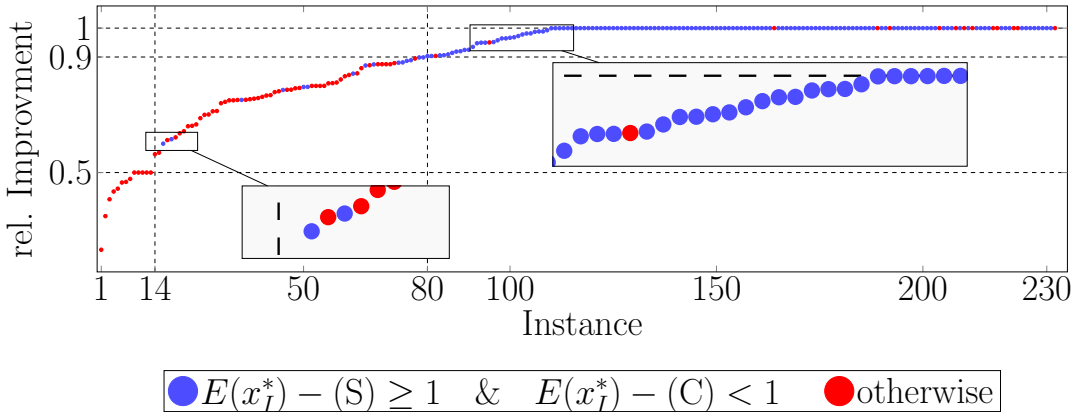


Figure 4.5: Comparison of lower bounds for (STD) and (CTG). We show relative improvement $^{(\mathrm{CTG})-(\mathrm{STD})}/_{E(x^*)-(\mathrm{STD})}$ on all problems, where we knew by either method the true optimal solution $E(x^*)$, but where (STD) is not tight ((STD) $< E(x^*)$). The lower bound (STD) was computed by **STD_relax**, while (CTG) was computed by **CTG_CB** and **CTG_relax**. A marker close to zero means no improvement, and close to one means our relaxation solved the instance exactly. We marked points with a blue circle if only (CTG) but not (STD) achieved a duality gap < 1, i.e., optimality, and used a red circle otherwise. For almost all instances, we have an improvement of 0.5 and for more than half of the instances improvement of 0.9.

CHAPTER 5

Integral Sparse and Cosparse Recovery

So far, we used convex priors to approximate sparse or cosparse signals, e.g., through the ℓ_1-norm or the anisotropic TV-seminorm. In Chapter 3, we employed the statistical dimension to relate the sparsity or cosparsity to the needed number of random linear measurements to uniquely reconstruct a sparse or cosparse signal by convex programming.

In this chapter, on the other hand, we consider non-convex sparse and cosparse signal recovery models. In the first part of this chapter, we focus on developing conditions for unique sparse or cosparse recovery. To this end, we employ the *union of subspace* model [LD08; BD09] as a low complexity model to relate the number of linear measurements for achieving unique recovery. Throughout this chapter, the analysis operator that defines the cosparse signal model is assumed to be the two-dimensional finite difference operator, also called image gradient operator. In this context, we show that the number of linear measurements required for unique reconstruction depends either on the number of jumps/discontinuities in the image gradient or on the number of connected components.

In subsequent sections 5.2.1 and 5.2.2, we investigate non-convex models for finding sparse or cosparse solutions subject to linear constraints. Given noiseless data, we consider the solution of

$$\begin{aligned} &\min \ \|\Omega x\|_0 \\ &\text{s.t. } Ax = b, x \in \mathcal{X}. \end{aligned} \tag{P_0}$$

Note that for $\Omega = \nabla$ the objective in (P_0) is also known as the *Potts prior* [Pot52; WS15].

In the presence of noisy data, we consider the cardinality constrained problem

$$\begin{aligned} &\min \ \|Ax - b\|_2^2 \\ &\text{s.t. } \|\Omega x\|_0 \leq s, x \in \mathcal{X}. \end{aligned} \tag{P_{ls}}$$

Since the problems (P_0) and (P_{ls}) are in general NP-hard [Bie96; Nat95], we cannot expect to find an efficient method to find the optimal solution.

For approximatively solving (P_0), one may consider the Lagrangian formulation, as done in [WS15; Sto15] for the Potts prior, i.e., $\Omega = \nabla$. These works, consider an ADMM splitting framework that leads to computationally tractable subproblems by dynamic programming. However, in general the obtained solution is not feasible for the linear system in (P_0). A greedy approach for solving directly (P_0) is proposed in [Nam13] with guaranteed performance only in the case of matrices A, Ω satisfying an extended version of the *exact recovery condition* (ERC) [Nam13, Cor. 8]. In the general case, there are no specialized approaches for approximating the cosparse *analysis* (P_0) and (P_{ls}).

Section 5.2 outlines two recent and promising approaches solving (P_0) and (P_{ls}) heuristically in case Ω is the identity matrix. They seem easily extendable to the general case, but failed in our experiments for every instance. Therefore, we only sketch the extension but omit any convergence proof or success rates guarantees.

Aiming at finding the global optimal solution to (P_0) and (P_{ls}), we reformulate these models as integer programs [Sch99; KV18; Bal09] such that we can apply standard methods as branch and bound [LD60; Dak65] in Section 5.3. Even though branch and bound finds a globally optimal solution in finite time, however, it may be necessary to enumerate all feasible solutions, which is not tractable in practice. Therefore, we compare different relaxations in case of the finite difference operator and develop specialized constraints to speed-up the branch and bound algorithm verifying optimality.

Using the finite difference operator, both models (P_0) and (P_{ls}) only target the number of jumps/discontinuities. In Section 5.4, we present linear constraints such that we limit the number of connected components in the reconstructed signal as motivated by the results derived in Section 5.1. Consequently, we obtain models needing less linear measurements to uniquely reconstruct an image with respect to the number of connected components.

5.1 Theoretical Recovery

In this section, we recall and extend the sampling requirements for classes of signals that live in a union of subspaces. The general results developed in Section 5.1.1 serve as a guideline for designing new recovery guarantees in Section 5.1.2 for the finite differences operator.

5.1.1 Union of Subspaces Model

Consider problem (P_0). We seek a signal $x \in \mathbb{R}^n$ corresponding to $\Omega_{\Lambda,\bullet}x = 0$ and $|\Lambda| = k$, where $\Lambda \subseteq [p]$. Defining $\Gamma_k := \{\Lambda \subseteq [p] \mid |\Lambda| \leq k\}$, a solution x^* with $\|\Omega x^*\|_0 = k$ to (P_0) is unique if and only if x^* is the only solution to

$$Ax = b \text{ s.t. } x \in \bigcup_{\Lambda \in \Gamma_k} \mathcal{N}(\Omega_{\Lambda,\bullet}). \tag{5.1}$$

Subsequently, we recall the results from [LD08; BD09] which show that for almost all matrices A, with a sufficient number of rows, (5.1) has at most one solution.

First, we define the invertibility of a *non-square* matrix A on some set $\mathcal{X}$.

Definition 5.1. Let $\mathcal{X} \subseteq \mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$. We say that A is invertible on $\mathcal{X}$ if for all $x, y \in \mathcal{X}$ it holds

$$A(x - y) = 0 \iff x = y. \tag{5.2}$$

The next proposition provides us necessary and sufficient condition for (P_0) to have at most one solution.

Proposition 5.2 ([LD08, Prop. 1])**.** *Let $A \in \mathbb{R}^{m \times n}, \Omega \in \mathbb{R}^{p \times n}$ and $\Gamma \subseteq 2^{[p]}$. Then it holds that*

$$\begin{gathered} A \text{ is invertible on } \bigcup_{\Lambda \in \Gamma} \mathcal{N}(\Omega_{\Lambda,\bullet}) \\ \iff \\ A \text{ is invertible on every } \mathcal{N}(\Omega_{\Lambda_1,\bullet}) + \mathcal{N}(\Omega_{\Lambda_2,\bullet}) \text{ with } \Lambda_1, \Lambda_2 \in \Gamma. \end{gathered}$$

Proof. See [LD08, Prop. 1]. □

Proposition 5.3. *Let $A \in \mathbb{R}^{m \times n}, \Omega \in \mathbb{R}^{p \times n}$ and $\Gamma \subseteq 2^{[p]}$. It holds for any $\Lambda_1, \Lambda_2 \in \Gamma$ that*

$$\begin{gathered} A \text{ is invertible on } \mathcal{N}(\Omega_{\Lambda_1,\bullet}) + \mathcal{N}(\Omega_{\Lambda_2,\bullet}) \\ \iff \\ \mathcal{N}(A) \cap (\mathcal{N}(\Omega_{\Lambda_1,\bullet}) + \mathcal{N}(\Omega_{\Lambda_2,\bullet})) = \{0\}. \end{gathered}$$

Proof. We start by defining $V = \mathcal{N}(\Omega_{\Lambda_1,\bullet}) + \mathcal{N}(\Omega_{\Lambda_2,\bullet})$. Since V is a subspace we have for any $x, y \in V$ that $x - y \in V$. As A is invertible on V then for any $A(x - y) = 0$ with $x, y \in V$ it follows that $x = y$. Consequently, $\mathcal{N}(A) \cap V = \{0\}$.

Conversely, we assume that $\mathcal{N}(A) \cap V = \{0\}$. Let $Ax = Ay$ with $x, y \in V$. Then $x - y \in V$ and $A(x - y) = 0$. If $x \neq y$ we would have found $0 \neq x - y \in \mathcal{N}(A) \cap V$ that violates our assumption. Therefore, we showed that $Ax = Ay$ implies $x = y$. This shows that A is invertible on V. □

The next proposition uses the conditions above to construct a lower bound on the number of rows of A such that (5.1) has at most one solution.

Proposition 5.4 ([LD08, Prop. 3]). *Let $A \in \mathbb{R}^{m\times n}, \Omega \in \mathbb{R}^{p\times n}$ and $\Gamma \subseteq 2^{[p]}$. If A is invertible on $\bigcup_{\Lambda\in\Gamma} \mathcal{N}(\Omega_{\Lambda,\bullet})$ it holds that*

$$m \geq \max_{\Lambda_1,\Lambda_2\in\Gamma} \dim \mathcal{N}(\Omega_{\Lambda_1,\bullet}) + \mathcal{N}(\Omega_{\Lambda_2,\bullet}). \tag{5.3}$$

Proof. See [LD08, Prop. 3]. □

The next theorem reveals that the lower bound (5.3) is necessary and sufficient to guarantee that (P_0) has at most one solution.

Theorem 5.5 ([BD09, Thm. 2.3]). *Let $A \in \mathbb{R}^{m\times n}, \Omega \in \mathbb{R}^{p\times n}$, $\Gamma \subseteq 2^{[p]}$ and m satisfying (5.3). Then for almost all A it holds that A is invertible on $\bigcup_{\Lambda\in\Gamma} \mathcal{N}(\Omega_{\Lambda,\bullet})$.*

Proof. See [BD09, Thm. 2.3]. □

Remark 5.1. In general (P_{ls}) does not yield a unique solution. Consider x^* as an optimal solution to (P_{ls}) and $\Lambda = \operatorname{cosupp}(\Omega x^*)$. If A is invertible on the feasible set of (P_{ls}) then x^* is unique with respect to $\Omega_{\Lambda,\bullet} x = 0$. Consequently, for almost all A with sufficiently many rows, (P_{ls}) has only a finite number of solutions.

Proposition 5.4 gives us a lower bound on the number of linear measurements such that A is invertible on a union of subspaces. Furthermore, Theorem 5.5 states that for almost all matrices the lower bound is sufficient. Now, we address the question whether a given A is invertible on some union of subspaces $\bigcup_{\Lambda\in\Gamma} \mathcal{N}(\Omega_{\Lambda,\bullet})$. In [Nam13] a property called *mutual independence* of two matrices is described. The authors in [Nam13] call two matrices A, Ω mutually independent if there are no non-trivial linear dependencies among the rows of $A \in \mathbb{R}^{m\times n}$ and $\Omega \in \mathbb{R}^{p\times n}$. In the next proposition, we show that the mutual independence property from [Nam13] is sufficient for the invertibility of A on a union of subspaces.

Proposition 5.6. *Let $A \in \mathbb{R}^{m\times n}, \Omega \in \mathbb{R}^{p\times n}$ with $p + m > n$. If for every set $\Theta \subseteq [m], \Lambda \subseteq [p]$ with $|\Theta| + |\Lambda| = n$ it holds that*

$$\mathcal{N}(A_{\Theta,\bullet}) \cap \mathcal{N}(\Omega_{\Lambda,\bullet}) = \{0\}, \tag{5.4}$$

then A is invertible on $\bigcup_{\Lambda\in\Gamma} \mathcal{N}(\Omega_{\Lambda,\bullet})$ with $\Gamma \subseteq 2^{[p]}$ if m satisfies Proposition 5.4.

Proof. Throughout the proof, we make use of the following equivalence

$$\mathcal{N}(A_{\Theta,\bullet}) \cap \mathcal{N}(\Omega_{\Lambda,\bullet}) = \{0\} \tag{5.5}$$

$$\Leftrightarrow \mathcal{R}(A_{\Theta,\bullet}^\top) + \mathcal{R}(\Omega_{\Lambda,\bullet}^\top) = \mathbb{R}^n. \tag{5.6}$$

In view of Proposition 5.2 and 5.3 we need to show that for every $\Lambda_1, \Lambda_2 \in \Gamma$ it follows that

$$\mathcal{N}(A) \cap (\mathcal{N}(\Omega_{\Lambda_1,\bullet}) + \mathcal{N}(\Omega_{\Lambda_2,\bullet})) = \{0\}. \tag{5.7}$$

Note that (5.7) above is equivalent to

$$\mathcal{R}(A^\top) + \mathcal{R}(\Omega^\top_{\Lambda_1,\bullet}) \cap \mathcal{R}(\Omega^\top_{\Lambda_2,\bullet}) = \mathbb{R}^n \,. \tag{5.8}$$

By contradiction, we assume that we found $\Lambda_1, \Lambda_2 \in \Gamma$ such that (5.8) fails. According to the mutual independence assumption (5.4) and Proposition 5.4, we have that $\dim \mathcal{R}(A^\top) = m$ and $\dim \mathcal{R}(\Omega^\top_{\Lambda_1,\bullet}) \cap \mathcal{R}(\Omega^\top_{\Lambda_2,\bullet}) \geq n - m$. If $|\Lambda_1| \geq n$ and $|\Lambda_2| \geq n$ we obtain from (5.4) that already $\mathcal{R}(\Omega^\top_{\Lambda,\bullet}) \cap \mathcal{R}(\Omega^\top_{\Lambda_2,\bullet}) = \mathbb{R}^n$. Consequently, $|\Lambda_1| < n$ or $|\Lambda_2| < n$. W.l.o.g. we assume that $|\Lambda_1| < n$.

As (5.8) does not hold, and $n - m \leq \dim \mathcal{R}(\Omega^\top_{\Lambda_1,\bullet}) \cap \mathcal{R}(\Omega^\top_{\Lambda_2,\bullet}) < n$, as shown above, there exist a row a_j with $j \in [m]$ in A such that

$$a_j^\top = \Omega^\top_{\Lambda_1,\bullet} \alpha \tag{5.9}$$

with $\alpha \in \mathbb{R}^{|\Lambda_1|}$. Next, we choose $\Theta' \subset [m]$ such that $|\Theta'| = n - |\Lambda_1| - 1$. Further, we define $\Theta = \Theta' \cup \{j\}$ and obtain $|\Theta| + |\Lambda_1| = n$. By construction we obtain

$$\mathcal{R}(A^\top_{\Theta,\bullet}) + \mathcal{R}(\Omega^\top_{\Lambda_1,\bullet}) \neq \mathbb{R}^n \,. \tag{5.10}$$

This, however, is a contradiction to (5.4) above. □

First, we introduce the following measure similar to [Nam13]

$$\kappa_\Omega(\Gamma) := \max_{\Lambda \in \Gamma} \dim \mathcal{N}(\Omega_{\Lambda,\bullet}), \qquad \Gamma \subseteq 2^{[p]}. \tag{5.11}$$

Next, we provide a sufficient condition for the invertibility of A on union of subspaces that involves the number of measurements m.

Proposition 5.7. *Let $A \in \mathbb{R}^{m\times n}, \Omega \in \mathbb{R}^{p\times n}$ and $\Gamma \subseteq 2^{[p]}$. If $\kappa_\Omega(\Gamma) \leq \frac{m}{2}$ then almost all A are invertible on $\bigcup_{\Lambda\in\Gamma} \mathcal{N}(\Omega_{\Lambda,\bullet})$.*

Proof. Since $\kappa_\Omega(\Gamma) \leq \frac{m}{2}$, than for every $\Lambda \in \Gamma$ we have that $\dim \mathcal{N}(\Omega_{\Lambda,\bullet}) \leq \frac{m}{2}$. Then it follows that

$$\dim \mathcal{N}(\Omega_{\Lambda_1,\bullet}) + \mathcal{N}(\Omega_{\Lambda_2,\bullet}) \leq m \tag{5.12}$$

for any $\Lambda_1, \Lambda_2 \in \Gamma$. Consequently, by Theorem 5.5 almost all $A \in \mathbb{R}^{m\times n}$ are invertible on $\bigcup_{\Lambda\in\Gamma} \mathcal{N}(\Omega_{\Lambda,\bullet})$. □

Remark 5.2. The authors from [Nam13, Prop. 3] derived the same conclusion as in Proposition 5.7. However, their proof differs from ours by including that A, Ω satisfy the mutual independence property (5.4).

Remark 5.3. Let $\Omega = I$ the identity and $\Gamma_s = \{\Lambda \subseteq [n] \mid |\Lambda| \geq n - s\}$. Then $\bigcup_{\Lambda\in\Gamma} \mathcal{N}(\Omega_{\Lambda,\bullet})$ contains all vectors $x \in \mathbb{R}^n$ with less than s nonzero entries, i.e., $\|x\|_0 \leq s$. The guarantees from Proposition 5.2 and 5.3 imply that $A \in \mathbb{R}^{m\times n}$ is invertible on $\bigcup_{\Lambda\in\Gamma} \mathcal{N}(\Omega_{\Lambda,\bullet})$ if and only if every choice of $k \leq 2s$ column vectors of A are in general position. Indeed, this is condition (2.28) that involves the spark to guarantee sparse recovery.

5.1.2 Finite Difference Operator

When choosing Ω in (P_0) or (P_{ls}) to be the two-dimensional finite difference operator, we are enforcing solutions (2D images) that exhibit constant regions and a minimal respectively bounded edge length. Prior analyzing the resulting union of subspaces induced by the finite difference operator, we shortly recall some definitions.

We identify an image by a grid graph $G = (\mathcal{V}, \mathcal{E})$, where each $v \in \mathcal{V}$ corresponds to a pixel. Let $(x_v)_{v\in\mathcal{V}} \in \mathbb{R}^{|\mathcal{V}|}$ be a vector. The *discrete gradient* of an image is defined by

$$\nabla x := (x_i - x_j)_{ij\in\mathcal{E}}. \tag{5.13}$$

Hence, we regard the finite difference operator $\nabla \in \{-1, 0, 1\}^{|\mathcal{E}|\times|\mathcal{V}|}$ as the matrix, where each row corresponds to one edge with exactly two nonzero entries. Further, we define for a subset $\Lambda \subseteq \mathcal{E}$ the following relation

$$v \sim_\Lambda w \Leftrightarrow \text{ there is a path in } \Lambda \text{ from } v \text{ to } w \text{ or } v = w. \tag{5.14}$$

For the definition of a path, we refer to Section 2.1.3. Clearly $\sim_\Lambda$ defines a equivalence relation on $\mathcal{V}$. It is reflexive by definition and symmetric as the graph is undirected. Transitivity follows by concatenating two paths.

Proposition 5.8. *Let $\Lambda \subseteq \mathcal{E}$ and $\mathcal{C}_\Lambda := \mathcal{V}/\sim_\Lambda$, then*

$$\mathcal{N}(\nabla_{\Lambda,\bullet}) = \operatorname{lin}\{\mathbb{1}_C \mid C \in \mathcal{C}_\Lambda\} \text{ and } \dim\mathcal{N}(\nabla_{\Lambda,\bullet}) = |\mathcal{C}_\Lambda|. \tag{5.15}$$

*The set $\mathcal{C}_\Lambda$ consists of all **connected components** with respect to Λ.*

Proof. Since $\sim_\Lambda$ is an equivalence relation, it follows that for every $C_1, C_2 \in \mathcal{C}_\Lambda$ it holds $C_1 \cap C_2 = \emptyset$ or $C_1 = C_2$. Therefore, if $\nabla_{\Lambda,\bullet} x = 0$, we directly obtain $x = \sum_{C\in\mathcal{C}_\Lambda} \alpha_C \mathbb{1}_C$ for suitable coefficients $\alpha_C \in \mathbb{R}$. Hence, we proved the claim. □

Using Proposition 5.8, we define the following family of index sets

$$\Gamma_k^c := \{\Lambda \subseteq \mathcal{E} \ : \ \dim\mathcal{N}(\nabla_{\Lambda,\bullet}) \leq k\}, \tag{5.16}$$

such that $\bigcup_{\Lambda\in\Gamma_k^c} \mathcal{N}(\nabla_{\Lambda,\bullet})$ defines all images with at most k connected components.

Corollary 5.9. *Almost all $A \in \mathbb{R}^{m\times n}$ with*

$$m \geq 2 \cdot k = \kappa_\nabla(\Gamma_k^c) \tag{5.17}$$

are invertible on $\bigcup_{\Lambda\in\Gamma_k^c} \mathcal{N}(\nabla_{\Lambda,\bullet})$.

Proof. Evaluate $\kappa_\nabla(\Gamma_k^c)$ defined by (5.11) yields k. By Proposition 5.7, it follows that almost all $A \in \mathbb{R}^{m\times n}$ are invertible on $\bigcup_{\Lambda\in\Gamma_k^c} \mathcal{N}(\nabla_{\Lambda,\bullet})$ if $\frac{m}{2} \geq \kappa_\nabla(\Gamma_k^c)$. Hence, the claim follows. □

In Section 5.4, we introduce a set of constraints which are added to (P_0) and (P_{ls}) such that any solution has at most k connected components.

Next, we define the family of index sets

$$\Gamma_k^s = \{\Lambda \subseteq \mathcal{E} \mid |\Lambda| \geq k\}, \tag{5.18}$$

such that $\bigcup_{\Lambda \in \Gamma_k^s} \mathcal{N}(\nabla_{\Lambda,\bullet})$ defines all images with a total edge length of at most $|\mathcal{E}| - k$. In the next proposition, we obtain a sharp bound on the individual nullspace.

Proposition 5.10 ([Nam13, Prop. 6])**.** *Let Ω be the finite difference operator and $|\mathcal{V}| = n$. Then for $k \geq 5$ we have*

$$n - \frac{k}{2} - \sqrt{\frac{k}{2}} - 1 \leq \kappa_\Omega(\Gamma_k^s) \leq n - \frac{k}{2} - \sqrt{\frac{k}{2}} + \frac{1}{2}. \tag{5.19}$$

Proof. The proof can be found in [Nam13]. □

Corollary 5.11. *Almost all $A \in \mathbb{R}^{m \times n}$ with*

$$m \geq 2 \cdot \left(n - \frac{k}{2} - \sqrt{\frac{k}{2}} + \frac{1}{2} \right) \geq 2 \cdot \kappa_\nabla(\Gamma_k^s) \tag{5.20}$$

are invertible on $\bigcup_{\Lambda \in \Gamma_k^s} \mathcal{N}(\nabla_{\Lambda,\bullet})$.

Proof. Plugging $\kappa_\Omega(\Gamma_k^s)$ from (5.19) into Proposition 5.7, hence the claim follows. □

Remark 5.4. In [Nam13] the authors additionally assume mutual independence in the Corollary 5.11. In view of Proposition 5.7, we do not need this assumption.

In summary, Corollaries 5.11 and 5.9 give us a lower bound on the needed number of linear measurements to uniquely reconstruct a signal by (P_0). As an example, if we have an image x^* with only two connected components, then for almost all matrices $A \in \mathbb{R}^{m \times n}$ with $m \geq 4$ and $b = Ax^*$ the model (P_0) has only x^* as a optimal solution.

Regarding (P_{ls}), the lower bound on the measurements implies a finite number of optimal solutions.

To conclude, the number of jumps or connected components accurately describes the minimal number of linear measurements necessary to reconstruct an image uniquely. While the number of jumps depends on the image discretization, the number of connected components is independent of the resolution.

Motivated by this theoretical findings we introduce, in Section 5.4, additional linear constraints to problems (P_0) and (P_{ls}) that limit the number of connected components.

5.2 Sparse Approximation

As previously mentioned, problems (P_0) and (P_{ls}) are NP-hard [Bie96; Nat95]. Hence, it is reasonable to assume that no efficient algorithm exists to solve them, unless NP=P.

Existing approaches for (P_0) proceed by applying the Lagrange dual, e.g., Potts regularized least-squares [WS15; Sto15], or developing greedy methods, e.g., [Nam13]. Without efficient primal heuristics, the Lagrange dual does not provide feasible solutions for the linear system. Success guarantees for greedy methods have strong assumptions on A, Ω like the extended version of the exact recovery condition (ERC) [Nam13, Cor. 8].

In this chapter, we mainly consider methods that find the optimal solution to (P_0) and (P_{ls}) without additional conditions to the matrices A, Ω. In particular, we examine recently published approximation algorithms [ZK15] and [Bra18] for $\Omega = I$ that are possibly generalizable to the analysis case. Unfortunately both methods were unable to solve any of our test instances when extending them directly to the analysis case. Therefore, we only outline our extension of these methods and do not provide convergence results or success guarantees. Further work is needed to incorporate the interesting ideas from [ZK15] and [Bra18] to the analysis models (P_0) and (P_{ls}).

In Section 5.2.1 below, we outline the approach from [ZK15] for (P_0) in case of $\Omega = I$. The authors from [ZK15] reformulate (P_0) into a *bilevel program* [Dem15, Chap. 2] and develop a parameterized convex program as an approximation. Subsequently, in subsection 5.2.2, we introduce the method in [Bra18] that uses a sequence of nonlinear non-convex programs to approximate a solution for (P_{ls}) in case of $\Omega = I$.

5.2.1 Weighted ℓ_1-Minimization for the Analysis Model

In the following, we discuss the work of [ZK15] which transforms (P_0) (for $\Omega = I$) into a bilevel program. A key ingredient to build the bilevel program is the relation between ℓ_0-minimization and weighted ℓ_1-minimization. First, we formulate the weighted ℓ_1-minimization problem (W_1) for our case:

$$\min \|W\Omega x\|_1 \text{ s.t. } Ax = b, \tag{W_1}$$

where $W = \operatorname{diag}(w), w \in \mathbb{R}^p_{>0}$. Next, we show a connection between (W_1) and (P_0). Let $x^* \in \mathbb{R}^n$, $\Lambda = \operatorname{cosupp}(\Omega x^*)$ and $\mathcal{N}(A) \cap \mathcal{N}(\Omega_{\Lambda,\bullet}) = \emptyset$. Then there are weights $w \in \mathbb{R}^p_{>0}$ for (W_1) such that x^* is the unique minimizer. Before we detail this result in Proposition 5.13 below, we collect some observations in the following lemma.

Lemma 5.12. *Let $W = \operatorname{diag}(w), w \in \mathbb{R}^p_{>0}, \Omega \in \mathbb{R}^{p\times n}$ and $x \in \mathbb{R}^n$. Then it holds*

(i) $\mathcal{N}(\Omega) = \mathcal{N}(W\Omega)$;
(ii) $\operatorname{sign}(\Omega x) = \operatorname{sign}(W\Omega x)$;
(iii) $\operatorname{cosupp}(\Omega x) = \operatorname{cosupp}(W\Omega x)$.

Proof. The items (ii) and (iii) are straightforward, as one set $y = \Omega x$ and use that $\operatorname{sign}(Wy)$ and $\operatorname{cosupp}(Wy)$ act element-wise without taking the magnitude of the element into account.

Denoting ω_i as the i-th row of Ω, we have

$$\begin{aligned} & x \in \mathcal{N}(\Omega) && && (5.21)\\ \Leftrightarrow\ & \langle \omega_i, x \rangle = 0 && \forall i \in [p] && (5.22)\\ \overset{w_i>0}{\Leftrightarrow}\ & w_i \cdot \langle \omega_i, x \rangle = 0 && \forall i \in [p] && (5.23)\\ \Leftrightarrow\ & x \in \mathcal{N}(W\Omega). && && (5.24) \end{aligned}$$

Observing that $x \in \mathcal{N}(\Omega) \Leftrightarrow x \in \mathcal{N}(W\Omega)$ completes the proof.

□

Proposition 5.13. *Let $x^* \in \mathbb{R}^n$ with $\Lambda = \operatorname{cosupp}(\Omega x)$ such that it holds*

$$\mathcal{N}(A) \cap \mathcal{N}(\Omega_{\Lambda,\bullet}) = \{0\}. \tag{5.25}$$

There exist weights $W = \operatorname{diag}(w)$ with $w \in \mathbb{R}^p_{>0}$ such that x^* is the unique minimizer of (W_1).

Proof. By Theorem 3.3, Lemma 5.12 and (5.25), we have that x^* is the unique minimizer if and only if

$$\exists \alpha \in \mathbb{R}^p : \Omega^\top W \alpha \in \mathcal{R}(A^\top), \alpha_{\Lambda^c} = \operatorname{sign}(\Omega_{\Lambda^c,\bullet} x^*), \|\alpha_\Lambda\|_\infty < 1. \tag{5.26}$$

Next, we discuss how to choose weights $w \in \mathbb{R}^p_{>0}$ such that there exists an $\tilde{\alpha}$ that satisfies (W_1). Consider the following linear program and its dual

$$\begin{aligned} &\max\ 0 \\ &\text{s.t. } \Omega_{\Lambda,\bullet}^\top \alpha + A^\top y = d \end{aligned} \qquad \text{(P)} \qquad\qquad \begin{aligned} &\min\ \langle d, v \rangle \\ &\text{s.t. } \begin{pmatrix} \Omega_{\Lambda,\bullet} \\ A \end{pmatrix} v = 0 \end{aligned} \qquad \text{(D)}$$

with $d = -\Omega_{\Lambda^c,\bullet}^\top \operatorname{sign}(\Omega_{\Lambda^c,\bullet} x^*)$. By (5.25) the dual program has only $v = 0$ as solution. A feasible dual program implies that the primal program is feasible as well [Van14, Chap. 5]. Consequently, it exists α^* that satisfies

$$\Omega_{\Lambda,\bullet}^\top \alpha^* + A^\top y = d. \tag{5.27}$$

Using α^*, we choose some $k > 1$ and define weights

$$w_\Lambda = k \cdot |\alpha^*| \text{ and } w_{\Lambda^c} = \mathbb{1}. \tag{5.28}$$

Considering (W_1) with the weights defined above, we construct a certificate $\tilde{\alpha}$ to show that x^* is the unique minimizer. Setting $\tilde{\alpha}_\Lambda = \frac{1}{k} \operatorname{sign}(\alpha^*)$ and $\tilde{\alpha}_{\Lambda^c} = \Omega_{\Lambda^c,\bullet}^\top \operatorname{sign}(\Omega_{\Lambda^c,\bullet} x^*)$, we calculate

$$\Omega^\top W \tilde{\alpha} = \Omega_{\Lambda,\bullet}^\top \alpha^* + \Omega_{\Lambda^c,\bullet}^\top \operatorname{sign}(\Omega_{\Lambda^c,\bullet} x^*) \in \mathcal{R}(A^\top). \tag{5.29}$$

Since $\|\tilde{\alpha}_\Lambda\|_\infty = \frac{1}{k} < 1$ holds by construction, it follows that x^* is the unique minimizer with the weights as in (5.28). □

Remark 5.5. If we drop the condition (5.25) in the Proposition 5.13, x^* may not the unique minimizer to (W_1). Let $\hat{x}$ be another minimizer of (W_1), then it holds $\operatorname{supp}(\Omega x^*) = \operatorname{supp}(\Omega \hat{x})$ and consequently $\|\Omega x^*\|_0 = \|\Omega \hat{x}\|_0$.

Remark 5.6. In view of Theorem 3.3 one expects that if we add box constraints to (W_1), we have to require

$$\mathcal{N}(A) \cap \mathcal{N}(\Omega_{\Lambda,\bullet}) \cap \mathcal{N}(\Psi) = \{0\}, \tag{5.30}$$

to achieve a similar result as in the previous Proposition 5.13. However, this is not true. Consider the following example

$$\begin{aligned} &\min \|x\|_1 \\ &\text{s.t. } \begin{pmatrix} 1 & 1 & 0 \end{pmatrix} x = \frac{1}{2} \\ &\qquad x \in [0,1]^3 \end{aligned} \tag{5.31}$$

with $x^* = \begin{pmatrix} 0 & \frac{1}{2} & 1 \end{pmatrix}^\top$ and $\Lambda = \operatorname{cosupp}(x^*)$. Clearly (5.30) holds, but there cannot exist any weights to make x^* the optimal solution to (W_1).

Corollary 5.14. *Let x^* be a unique solution to (P_0). There exist weights $w \in \mathbb{R}^n_{>0}$ such that x^* is the unique solution to (W_1).*

Proof. Since x^* is a unique solution to (P_0), and $\mathcal{N}(A) \cap \mathcal{N}(\Omega_{\Lambda,\bullet}) = \{0\}$ is a necessary condition such that x^* is a unique solution, the claim follows directly by Proposition 5.13. □

So far we have seen that (W_1) is used to reconstruct a sparse signal with respect to Ω for an appropriate choice of weights $w \in \mathbb{R}^n_{>0}$. Next, we discuss how to find these weights without knowing the unique optimal solution of (W_1) in advance.

In the following, we rephrase (W_1) to a linear program (PW_1) and formulate the corresponding dual program (DW_1).

$$\begin{aligned} &\min \langle t, w \rangle \\ &\text{s.t. } Ax = b, \\ &\qquad \Omega x + t - \alpha = 0, \\ &\qquad -\Omega x + t - \beta = 0, \\ &\qquad \alpha, \beta, t \geq 0. \end{aligned} \tag{PW_1}$$

$$\begin{aligned} &\max \langle b, y \rangle \\ &\text{s.t. } A^\top y + \Omega^\top (u - v) = 0, \\ &\qquad u + v + s = w, \\ &\qquad u, v, s \geq 0. \end{aligned} \tag{DW_1}$$

By strict complementarity [Van14, Thm. 10.7], we obtain the existence of a primal solution pair t^*, x^* of (PW_1) and a dual solution s^* of (DW_1) such that $t^* + s^* > 0$ and $t_i^* \cdot s_i^* = 0$ for all $i \in [p]$. Since $\|t^*\|_0 = \|\Omega x^*\|_0$, we get $\|\Omega x^*\|_0 + \|s^*\|_0 = p$. As a consequence, the sparsity of Ωx^* in the primal program corresponds to the density of s^* in the dual program.

Based on the above observation, we proceed as in [ZK15] and construct the following bilevel program [Dem15]

$$\begin{aligned} \max\ & \|s\|_0 \\ \text{s.t. } & A^\top y = \Omega^\top (u - v), \\ & u + v + s = w, \\ & \langle b, y \rangle = \min\{\|W\Omega x\|_1 \mid Ax = b\}, \\ & u, v, s \geq 0, \end{aligned} \tag{BP}$$

where $W = \operatorname{diag}(w)$.

Proposition 5.15. *Let x^* be the unique minimizer of (P_0). Then (BP) calculates the optimal weights $\tilde{w}$ such that x^* is the unique minimizer of (W_1).*

Proof. By Proposition 5.13, it follows that there are weights $w \in \mathbb{R}^n_{>0}$ such that x^* is the unique minimizer of (W_1). Consequently, $t^* = \Omega x^*$ is the unique minimizer of (PW_1). By strict complementarity, there exists a dual solution s^* of (DW_1) such that $\|s^*\|_0 = p - \|\Omega x^*\|_0$ holds. Therefore, the optimal weights $\tilde{w}$ belong to the feasible set of (BP).

Assuming that there exists $\hat{s}$ in the feasible set of (BP) with $\|\hat{s}\|_0 > \|s^*\|_0$. Then $\hat{s}$ is part of a dual solution corresponding to $\hat{w}$ from (DW_1). Complementary slackness then yields for all primal solutions $\hat{t}$ that $\hat{t}_i \cdot \hat{s}_i = 0$ for all $i \in [p]$. This implies that for any optimal solution $\hat{x}$ to (W_1) with $\hat{w}$ it follows $\|\Omega x^*\|_0 > \|\Omega \hat{x}\|_0$. This is a contradiction to the optimality of x^*. Additionally, if $\|\hat{s}\|_0 = \|s^*\|_0$, again we obtain $\hat{x}$ from (W_1) with $\hat{w}$ and it follows $\|\Omega x^*\|_0 = \|\Omega \hat{x}\|_0$. Since we assumed that x^* is the unique minimizer, it follows $x^* = \hat{x}$. □

In general, bilevel programs are hard to solve and there are no generic algorithms. Therefore, [ZK15] proposed a relaxation to approximate the solution of (BP). First, we observe that scaling the weights for (W_1) does not change the solution. Having weights w, we get the same solutions when choosing $\lambda \cdot w$ for $\lambda > 0$. For that reason, we assume that

$$1 = \langle b, y \rangle = \min\{\|W\Omega x\|_1 \mid Ax = b\}. \tag{5.32}$$

Before we write down the relaxation of (BP), we need the following class of functions.

Definition 5.16 ([ZK15])**.** We call a function $\Phi_\varepsilon : \mathbb{R}^p_{\geq 0} \to \mathbb{R}$ with $\varepsilon > 0$ a merit function if it holds that

(i) for any $s \in \mathbb{R}^p_{\geq 0}$ it holds $\Phi_\varepsilon(s) \to \|s\|_0$ if $\varepsilon \to 0$;

(ii) Φ_ε is continuously differentiable and concave with respect to s over an open set containing $\mathbb{R}^p_{\geq 0}$;

(iii) for any given constants $0 < c_1 < c_2$ there exists a small $\varepsilon^* > 0$ such that for any given $\varepsilon \in (0, \varepsilon^*]$ it holds

$$\Phi_\varepsilon(s) - \Phi_\varepsilon(\hat{s}) \geq \frac{1}{2}, \tag{5.33}$$

for any $0 \leq s, \hat{s} \leq c_2$ satisfying $\|\hat{s}\|_0 < \|s\|_0$ and $c_1 \leq s_i \leq c_2$ for all $i \in \operatorname{supp}(s)$.

Proposition 5.17 ([ZK15, Prop. 3.2])**.** *Let $\varepsilon \in (0, 1)$. All the following functions are merit function according to the Definition 5.16:*

$$\Phi_\varepsilon(s) = \sum_{i=1}^{n} \left(1 - e^{-\frac{s_i}{\varepsilon}}\right), \quad \text{where } s \in \mathbb{R}^n; \tag{5.34}$$

$$\Phi_\varepsilon(s) = \sum_{i=1}^{n} \frac{s_i}{s_i + \varepsilon}, \quad \text{where } s_i > -\varepsilon \text{ for all } i = 1, \ldots, n; \tag{5.35}$$

$$\Phi_\varepsilon(s) = n - \frac{1}{\log \varepsilon} \left(\sum_{i=1}^{n} \log\left(s_i + \varepsilon\right) \right), \text{ where } s_i > -\varepsilon \text{ for all } i = 1, \ldots, n. \tag{5.36}$$

Using the merit function from Definition 5.16, we now adapt the relaxation from [ZK15]

$$\begin{aligned} \max\ & \alpha \Phi_\varepsilon(s) + \langle b, y \rangle \\ \text{s.t } & A^\top y = \Omega^\top (u - v), \\ & s = w - u - v, \\ & \langle b, y \rangle \leq 1, \\ & w \in \mathcal{W}^k, 0 \leq s, u, v. \end{aligned} \tag{$\mathrm{BP}^{\alpha,\varepsilon,k}$}$$

The problem ($\mathrm{BP}^{\alpha,\varepsilon,k}$) above is considered as a perturbed version of (DW_1), where $\langle b, y \rangle \leq 1$ models the weak duality. However, if one sets $w \geq 0$ in (DW_1), then $s > 0$ (that yields $\|s\|_0 = p$) would always be a feasible solution. Therefore, the authors propose Algorithm 5.1 which iteratively adjusts the set $\mathcal{W}^k$ such that $\|s\|_0 = p$ is not feasible for ($\mathrm{BP}^{\alpha,\varepsilon,k}$).

As already mentioned, Algorithm 5.1 could not solve any of our test instances. For this reason, we refrain from further analysis, since translating success guarantees would be associated with very restrictive conditions to the matrices A, Ω. Concerning the application in discrete tomography, this is not target-oriented. However, the reformulation of (P_0) as a bilevel program is a interesting approach so that a more profound analysis might be interesting for other applications.

Algorithm 5.1: Bilevel Approximation [ZK15] for (P_0)

Input: $\alpha^0, \tau, \varepsilon, \alpha^*, \vartheta \in (0,1)$

Let x^0 be the solution of (W_1) with $W = I$
Choose $\Gamma^0 \geq \vartheta \max\left\{1, \frac{1}{\|x\|_1}\right\}$
Set $\mathcal{W}^0 = \{w \in \mathbb{R}^p_{\geq 0} \mid \langle |\Omega x^0|, w\rangle \leq \vartheta, w \leq \Gamma^0\}$
Set $k = 0$
while $\alpha^k > \alpha^*$ **do**
  Solve $(\mathrm{BP}^{\alpha,\varepsilon,k})$ and obtain w^k
  Set $W^k = \mathrm{diag}(w^k)$
  Solve (W_1) using W^k and obtain x^k
  Choose $\Gamma^{k+1} \geq \vartheta \max\left\{1, \frac{\|w^k\|_\infty}{\|W^k \Omega x^k\|_1}\right\}$
  Set $\mathcal{W}^{k+1} = \{w \in \mathbb{R}^p_{\geq 0} \mid \langle |\Omega x^k|, w\rangle \leq \vartheta, w \leq \Gamma^{k+1}\}$
  Set $\alpha^{k+1} = \tau \cdot \alpha^k$
  Set $k = k + 1$
end

5.2.2 Cardinality Constriants

In this subsection, we extend the method from [Bra18] to the analysis case of problem $(\mathrm{P}_{\mathrm{ls}})$. Our focus here is on the computational steps of the algorithm and not on its convergence theory.

In order to find a feasible (not necessary optimal) solution to $(\mathrm{P}_{\mathrm{ls}})$ for the case $\Omega = I$, the authors of [Bra18] propose iterations based on a sequence of non-linear, non-convex optimization problems. In particular, they formulate the following non-linear program

$$\begin{aligned} \min_{x,y}\ & f(x) \\ \text{s.t. }\ & g(x) \leq 0, h(x) = 0, \\ & 0 \leq y \leq \mathbb{1}, \\ & x \circ y = 0, \\ & \langle \mathbb{1}, y\rangle \geq n - s, \end{aligned} \tag{5.37}$$

where $f : \mathbb{R}^n \to \mathbb{R}, g : \mathbb{R}^n \to \mathbb{R}^r, h : \mathbb{R}^n \to \mathbb{R}^q$ are continuously differentiable functions. The problem (5.37) is an instance of a mathematical program with complementarity constraints (MPCC) [LPR96].

It is straightforward to see that for any feasible solution of the above problem $\|y\|_0 \geq n - s$ implies $\|x\|_0 \leq s$. Consequently, any optimal solution x^* to the above problem is also optimal for $(\mathrm{P}_{\mathrm{ls}})$ (when $\Omega = I$). Next, we modify the problem above to obtain

the general case

$$\begin{aligned} \min_{x,y,z} \; & f(x) \\ \text{s.t.} \; & g(x) \leq 0, h(x) = 0, \\ & \Omega x = z, \\ & 0 \leq y \leq \mathbb{1}, \\ & z \circ y = 0, \\ & \langle \mathbb{1}, y \rangle \geq n - s. \end{aligned} \tag{NLP}$$

Again, we get $\|y\|_0 \geq n - s$ implying $\|\Omega x\|_0 = \|z\|_0 \leq s$. Therefore, any optimal solution x^* of (NLP) is also optimal for (P_{ls}) and vice versa.

Since (NLP) is a non-convex optimization problem we can only expect to find local minima. In [Bra18] the authors use the Scholtes regularization [Sch01] to be less dependent on initialization and to find a „better “ of (NLP).

$$\begin{aligned} \min_{x,y,z} \; & f(x) \\ \text{s.t.} \; & g(x) \leq 0, h(x) = 0, \\ & \Omega x = z, \\ & 0 \leq y \leq \mathbb{1}, \\ & -t \cdot \mathbb{1} \leq z \circ y \leq t \cdot \mathbb{1}, \\ & \langle \mathbb{1}, y \rangle \geq n - s. \end{aligned} \tag{NLP^t}$$

Using the Scholtes regularization [Sch01], we now replace $z \circ y = 0$ in (NLP) through $-t \cdot \mathbb{1} \leq z \circ y \leq t \cdot \mathbb{1}$. Then, one solves (NLP^t) for a decreasing sequence $t^k \to 0$ and use any solution y^k, x^k, z^k as an initial solution for the next iteration $k + 1$.

As in [Bra18], we used the software SNOPT [GMS05] to solve the subproblems (NLP^t). However, we always encountered numerical problems if $z \circ y \approx 0$ such that we never reached a feasible solution for none of our test instances.

5.3 Exact Recovery by Integer Programming

In contrast to the previous section, we now discuss a way to provably find the optimal solution to (P_0) and (P_{ls}). To this end, we formulate our considered problems into *mixed-integer-programs* (MIP)[Sch99; KV18; Bal09]. Solving MIP might be extremely time-consuming. However, there are powerful commercial solver like GUROBI[1] or CPLEX[2] which handles large MIPs with millions of variables. By default, the mentioned software uses *branch and bound* to solve MIPs, which we introduced in Section 2.1.4.

In this section, we transform (P_0) and (P_{ls}) into a MIP. There are already for-

[1] https://www.gurobi.com/
[2] https://www.ibm.com/de-de/analytics/cplex-optimizer

mulations, e.g., [Bou16; JP07; Bie96] for $\Omega = I$ which are directly extendable for the general analysis case. Our analysis in Section 5.3.1 differs as it uses *disjunctive programming* [Bal18]. This leads us to an alternative MIP formulation that we now present.

We start by introducing decision variables $\alpha \in \{0,1\}^p$ and the following implications with respect to the considered problems in this chapter

$$\forall i \in [p] : \alpha_i = 0 \to (\Omega x)_i = 0. \tag{5.38}$$

It is straightforward to see that if $\alpha \in \{0,1\}^p$ and $x \in \mathbb{R}^n$ satisfies (5.38), then it holds

$$\|\Omega x\|_0 \leq \langle \alpha, \mathbb{1} \rangle. \tag{5.39}$$

For the later use, it is important to stress out that for a given $x \in \mathbb{R}^n$ it always exist $\alpha \in \{0,1\}^p$ satisfying (5.38) such that

$$\|\Omega x\|_0 = \langle \alpha, \mathbb{1} \rangle. \tag{5.40}$$

Now, we use (5.38) to transform (P_0) into

$$\begin{aligned} \min\ & \langle \alpha, \mathbb{1} \rangle \\ \text{s.t. } & Ax = b, \\ & \alpha_i = 0 \to (\Omega x)_i = 0, \qquad \forall i \in [p], \\ & \alpha \in \{0,1\}^p, \end{aligned} \tag{P_0^I}$$

and (P_{ls}) into

$$\begin{aligned} \min\ & \|Ax - b\|_2^2 \\ \text{s.t. } & \langle \alpha, \mathbb{1} \rangle \leq s, \\ & \alpha_i = 0 \to (\Omega x)_i = 0, \qquad \forall i \in [p], \\ & \alpha \in \{0,1\}^p. \end{aligned} \tag{$\mathrm{P}_{\mathrm{ls}}^I$}$$

The problems above belong to the class of optimizing problems with indicator constraints [Bon15; Bel16; FHS16].

After discussing our disjunctive programming approach in Section 5.3.1, we close with Section 5.3.2 by discussing additional constraints to reduce the depth of the resulting branch and bound tree when solving (P_0^I) and ($\mathrm{P}_{\mathrm{ls}}^I$).

5.3.1 Disjunctive Programming

In the field of (linear) disjunctive programming [Bal18] one deals with optimization problems, where the feasible set is a union of polyhedra. A polyhedra P is represented by a matrix $A \in \mathbb{R}^{m \times n}$ and right-hand side $b \in \mathbb{R}^m$ by

$$P = \{x \mid Ax \geq b\}. \tag{5.41}$$

Having some finite index set $\mathcal{I}$ we write the union of polyhedra as

$$x \in F := \bigcup_{k \in \mathcal{I}} \{y \mid A^k y \geq b^k\}. \tag{5.42}$$

The reason for the name „disjunctive“ comes from the fact that

$$x \in F \iff x \text{ statisfies } \bigvee_{k \in \mathcal{I}} A^k x \geq b^k. \tag{5.43}$$

Next, we translate (5.38) into the setting of *disjunctive programming*. Let $\alpha \in \mathbb{R}^p, x \in \mathbb{R}^n$ satisfying

$$\forall i \in [p] : (\alpha_i = 0 \wedge (\Omega x)_i = 0) \vee \alpha_i = 1. \tag{5.44}$$

Then $\alpha \in \{0,1\}^p$ and the pair x, α satisfy (5.38). Obviously, the reverse direction also applies. As a result, we get for each implication ($i \in [p]$) in (5.38) a union

$$\{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R}^p \mid \alpha_i = 0 \wedge (\Omega x)_i = 0\} \cup \{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R}^p \mid \alpha_i = 1\}. \tag{5.45}$$

Next, we present a result that gives us a linear representation of the convex hull for any union of polyhedra.

Theorem 5.18 ([Bal18, Thm. 2.1])**.** *The closed convex hull of F (defined as in (5.42)), $\overline{\text{conv}}(F)$, is the set of those x for which there exist $(y^k, y_0^k) \in \mathbb{R}^{n+1}, k \in \mathcal{I}$, satisfying*

$$\begin{aligned} x &= \sum_{k \in \mathcal{I}} y^k \\ A^k y^k - b^k y_0^k &\geq 0 \quad k \in \mathcal{I} \\ \sum_{k \in \mathcal{I}} y_0^k &= 1 \\ y_0^k &\geq 0 \qquad\quad k \in \mathcal{I}. \end{aligned} \tag{5.46}$$

In the sense of model complexity, one asks whether there is a representation with less variables as in Theorem 5.18. The answer is already given in [CDF19]. We summarize the corresponding result in the following theorem.

Theorem 5.19. *The representation given by Theorem 5.18 is optimal.*

Proof. See [CDF19]. □

Applying Theorem 5.18 to (5.44), we get the following linear representation of the convex hull

$$\begin{aligned} x &= y^i + z^i, && i \in [p], \\ (\Omega y^i)_i &= 0, && i \in [p], \\ y^i, z^i &\in \mathbb{R}^n, \alpha \in [0,1]^p. \end{aligned}$$

The linear system above is the convex hull (also a convex relaxation) to the indicator constraint in (P_0^I) and $(\mathrm{P}_{\mathrm{ls}}^I)$. Consequently, the relaxation above is useless when applying branch and bound directly. Hence, we assume $x \in [l,u]^n$ for $l < u$ and modify (5.44) to

$$\forall i \in [p] : (\alpha_i = 0 \wedge (\Omega x)_i = 0 \wedge x \in [l,u]^n) \vee (\alpha_i = 1 \wedge x \in [l,u]^n). \tag{5.47}$$

Now we apply again Theorem 5.18 and get the following linear system as the convex hull

$$\begin{aligned} &x = y^i + z^i, && i \in [p], \\ &(\Omega y^i)_i = 0, && i \in [p], \\ &y^i \in [(1-\alpha_i)\cdot l, (1-\alpha_i)\cdot u], && i \in [p], \\ &z^i \in [\alpha_i \cdot l, \alpha_i \cdot u], && i \in [p], \\ &\alpha \in [0,1]^p. \end{aligned} \tag{5.48}$$

Remark 5.7. In practical applications like discrete tomography one always may bound the values in the reconstructed image.

A more common approach in the mixed integer programming literature is called the *big-M* method [JP07; Bou16; Bon15] that considers (5.47) by

$$\begin{aligned} &-\alpha\theta \leq \Omega x \leq \alpha\theta, \\ &\alpha \in [0,1]^p, \\ &x \in [l,u]^n, \end{aligned} \tag{5.49}$$

where $\theta > 0$ is assumed to be big enough. This approach also yields a convex relaxation of (5.47). Considering the results above, we conclude that big-M method cannot yield a better convex relaxation then (5.48). Since (5.48) is independent of the choice of $\theta > 0$, we believe that this approach is preferable to the big-M method if no optimal $\theta > 0$ is known.

As an example, we consider $\Omega = \nabla$ as the two-dimensional finite difference operator as in Section 5.1.2. We use the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ defining the pairwise differences to write (5.47) as

$$\forall ij \in \mathcal{E} : (\alpha_{ij} = 0 \wedge x_i = x_j \wedge x_i, x_j \in [l,u]) \vee (\alpha_{ij} = 1 \wedge x_i, x_j \in [l,u]). \tag{5.50}$$

Analogously, we obtain, through (5.48), for each edge $ij \in \mathcal{E}$ the following linear system

$$\begin{aligned} &x_i = y^{ij} + z^{ij,i}, \\ &x_j = y^{ij} + z^{ij,j}, \\ &y^{ij} \in [(1-\alpha_{ij})\cdot l, (1-\alpha_{ij})\cdot u], \\ &z^{ij,i}, z^{ij,j} \in [\alpha_{ij}\cdot l, \alpha_{ij}\cdot u], \\ &\alpha_{ij} \in [0,1]. \end{aligned} \tag{5.51}$$

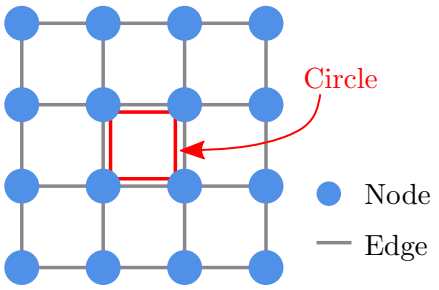


Figure 5.1: In this figure, we illustrate a simple grid graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and marked a circle in red.

As a consequence, both the big-M and the disjunctive programming approach above need $O(|\mathcal{E}|)$ many variables and constraints to construct a convex relaxation to (P_0^I) and $(\mathrm{P}_{\mathrm{ls}}^I)$ with respect to the finite difference operator.

5.3.2 Circle Constraints

When using branch and bound to solve (P_0^I) or $(\mathrm{P}_{\mathrm{ls}}^I)$, at each iteration a convex relaxation is solved (see Section 2.1.4). In particular, if the solution $\alpha^* \notin \{0,1\}^p$, one index $i \in [p]$ is chosen to create *two* subproblems with $\alpha_i = 0$ respectively $\alpha_i = 1$. As this procedure enumerates many possible choices of $\alpha \in \{0,1\}^p$, we should avoid to create unnecessary subproblems.

We now consider the case $\Omega = \nabla$. As in Section 5.1.2, we consider the corresponding graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which models the relation between the pixel. Let $C \subseteq \mathcal{E}$ be a circle in the graph $\mathcal{G}$ as in Figure 5.1. For the mathematical definition of a circle in a graph, we refer to Section 2.1.3. We then add the following constraints to (P_0^I) respectively $(\mathrm{P}_{\mathrm{ls}}^I)$

$$\alpha_{e'} \leq \sum_{e \in C \setminus \{e'\}} \alpha_e, \qquad \forall e' \in C, \tag{5.52}$$

without changing the feasible set. Suppose that we run branch and bound for some iterations and there is a node such that $\alpha_e = 0$ is fixed for all $e \in C \setminus \{e', e''\}$ with $e', e'' \in C$. Then (5.52) implies that $\alpha_{e'} = \alpha_{e''}$. Consequently, only two more subproblems are needed to decide whether $\alpha_{e'} = \alpha_{e''} = 1$ or $\alpha_{e'} = \alpha_{e''} = 0$. In Figure 5.2, we illustrate the use of circle constraints on a small example.

Adding all possible circles would result in an exponential number of constraints. In order to restrict this number, we only considered circles of length four.

In Figure 5.3, we show an experiment that illustrates the benefit of using circle constraints. We need to stress out, however, that circle constraints are only useful when proving optimality of an integer solution through branch and bound. We noticed in our experiments that when circle constraints, the time to find the best integer solution was increased dramatically. Hence, circle constraints only help to increase the

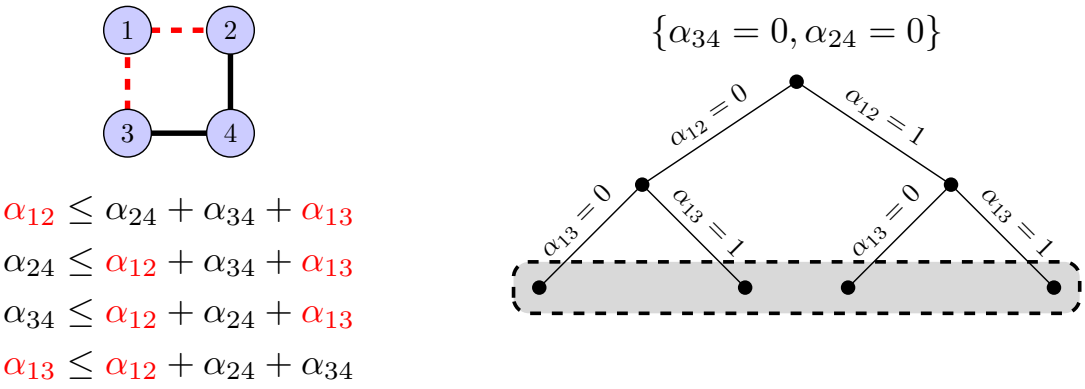


Figure 5.2: As an example, suppose we have four variables corresponding to the four grid points. Assuming that these four variables lie on a circle of the underlying grid graph, we obtain the circle constraints in the lower left. In our example, two edge variables are already fixed to zero (red dashed line), i.e., $\alpha_{12} = \alpha_{13} = 0$. Within a branch and bound run, six more subproblems (see right) would have to be enumerated to verify all possibilities for the last two edge variables. However, the circular constraints prevent branch and bound to consider of the four gray-shaded nodes, in view of $\alpha_{12} = \alpha_{13}$.

lower bound faster when testing optimality in branch and bound.

5.4 Connected Components Constraints

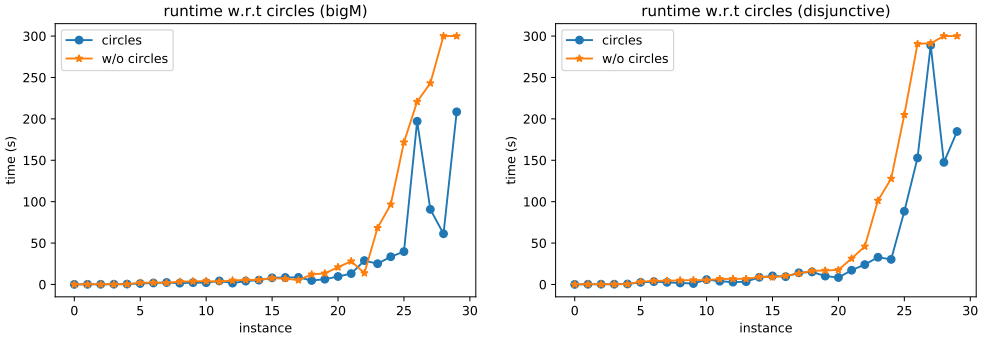
So far, we have focused on problems (P_0) and $(\mathrm{P}_{\mathrm{ls}})$. In view of $\Omega = \nabla$, these problems target the number of discontinuities. In view of Proposition 5.9, we derived in Section 5.1.2 that the number of connected components in an image are crucial to predict unique reconstruction by (P_0) and $(\mathrm{P}_{\mathrm{ls}})$. For that reason, we introduce in this section modifications of problems (P_0^I) and $(\mathrm{P}_{\mathrm{ls}}^I)$ that allow integer solutions with at most c connected components.

Consider now the finite difference operator and the corresponding grid graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Let $x \in \mathbb{R}^{|\mathcal{V}|}$ and $\alpha \in \{0,1\}^{|\mathcal{E}|}$ satisfying (5.50), i.e, each $\alpha_{ij} = 0$ for $ij \in \mathcal{E}$ implies $x_i = x_j$. Further, we define

$$\mathcal{G}(\alpha) = (\mathcal{V}, \mathcal{E}_\alpha) \tag{5.53}$$

where $\mathcal{E}_\alpha = \{e \in \mathcal{E} \mid \alpha_e = 0\}$. Hence, if x corresponds to a two-dimensional image, the connected components of $\mathcal{G}(\alpha)$ correspond to the constant regions of x. As a result, to restrict the number of connected components in the solutions of (P_0^I) and $(\mathrm{P}_{\mathrm{ls}}^I)$, it is enough to check whether the graph induced by α has at most c connected components.

In the following, we adapt the results from [Hoj18]. The authors in [Hoj18] used linear constraints from *network flow* [KV18, Chap. 8] to check if two nodes belong to



(a) 60% instances solved faster (b) 75% instances solved faster

Figure 5.3: For 30 images with four connected components and exactly 36 jumps, we reconstructed each image (using two binary tomography projections) through (P_0^I) (using GUROBI with timelimit of 300 seconds). The percentages in the captions indicate that the use of the circle constraints led to a reduction of the runtime in % percent of the considered instances.

the same connected component. Therefore, we introduce additional variables

$$f_{i\to j}, f_{j\to i} \in \mathbb{R}_{\geq 0} \qquad \forall ij \in \mathcal{E}, \tag{5.54}$$

where $f_{i\to j}$ models the *flow* from node i to j.

Next, we choose a subset $\mathcal{S} \subseteq \mathcal{V}$ and call each $s \in \mathcal{S}$ a *sink*. This then is used to define that

$$\alpha \in \mathcal{F}(\mathcal{S}) \tag{5.55}$$

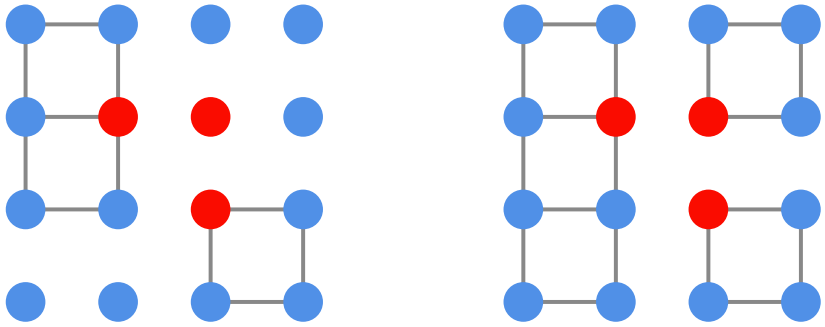
if and only if α satisfies the linear system

$$f_{i\to j} + f_{j\to i} \leq (|\mathcal{V}| + 1) \cdot (1 - \alpha_{ij}), \qquad ij \in \mathcal{E}, \tag{5.56a}$$

$$\sum_{ij\in\mathcal{E}} f_{i\to j} - f_{j\to i} \geq 1, \qquad i \in \mathcal{V} \setminus \mathcal{S}, \tag{5.56b}$$

$$f_{i\to j}, f_{j\to i} \geq 0, \qquad ij \in \mathcal{E}. \tag{5.56c}$$

Basically, (5.56) checks if for every node $v \in \mathcal{V} \setminus \mathcal{S}$ there is a path in $\mathcal{G}(\alpha)$ (see (5.53)) to any node of $\mathcal{S}$. By (5.56a), we ensure that the only nonzero flow variables are between nodes, which are adjacent with respect to $\mathcal{G}(\alpha)$. Except nodes contained in $\mathcal{S}$, for each $i \in \mathcal{V} \setminus \mathcal{S}$ there exist a node $j \in \mathcal{V}$ so that it holds $f_{i\to j} > 0$ or $f_{i\to j} > 0$. As a consequence, if $\alpha \in \mathcal{F}(\mathcal{S})$ then there exist for each $i \in \mathcal{V} \setminus \mathcal{S}$ a path in $\mathcal{G}(\alpha)$ to one of the nodes defined in $\mathcal{S}$. Hence, if $|\mathcal{S}| \leq c$, $\mathcal{G}(\alpha)$ has at most c connected components. We depict the $\mathcal{G}(\alpha)$ for $\alpha \in \mathcal{F}(\mathcal{S})$ and $\alpha \notin \mathcal{F}(\mathcal{S})$ in Figure 5.4. In case of $\alpha \notin \mathcal{F}(\mathcal{S})$ it need to exist a path from every non-sink node to any sink. If $\alpha \notin \mathcal{F}(\mathcal{S})$, there are nodes which separated from every sink.



(a) (b)

Figure 5.4: Both figures show a graph $\mathcal{G}(\alpha)$ for $\alpha \in \{0,1\}^{|\mathcal{E}|}$ with nodes (circle) and edges (line). The sinks defined by $\mathcal{S}$ are marked in red. Each edge depicted implies $\alpha_e = 0$ (otherwise $\alpha_e = 1$). For (a) it follows $\alpha \notin \mathcal{F}(\mathcal{S})$, as there are nodes with no path to any sink. On the other hand, for (b) it follows $\alpha \in \mathcal{F}(\mathcal{S})$, as for every node it exists a path to a red sink.

Now, we use (5.55) to reformulate (P_0^I) into

$$\begin{aligned} \min\ & \langle \alpha, \mathbb{1} \rangle \\ \text{s.t. } & Ax = b, \\ & \alpha_i = 0 \to (\Omega x)_i = 0 \qquad \forall i \in [p], \\ & \alpha \in \mathcal{F}(\mathcal{S}) \\ & \alpha \in \{0,1\}^p, \end{aligned} \tag{P_0^c}$$

and $(\mathrm{P}_{\mathrm{ls}}^I)$ into

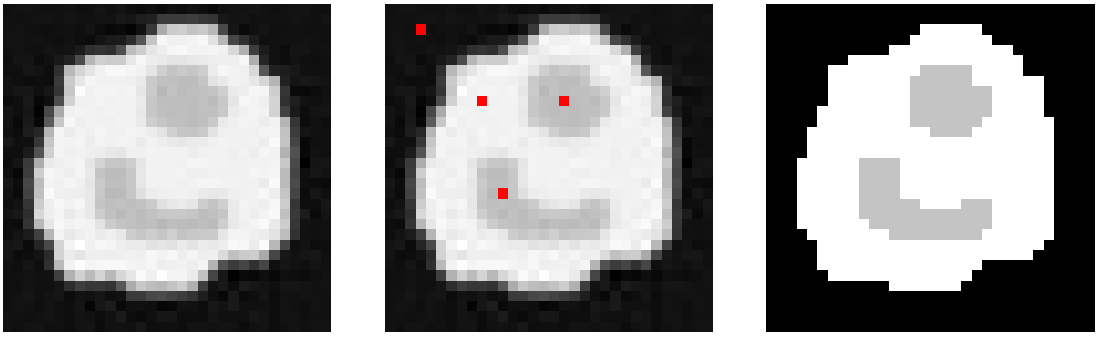
$$\begin{aligned} \min\ & \|Ax - b\|_2^2 \\ \text{s.t. } & \alpha \in \mathcal{F}(\mathcal{S}), \\ & \alpha_i = 0 \to (\Omega x)_i = 0 \qquad \forall i \in [p], \\ & \alpha \in \{0,1\}^p. \end{aligned} \tag{$\mathrm{P}_{\mathrm{ls}}^c$}$$

Hence, if $|\mathcal{S}| \leq c$ every solution to (P_0^c) or $(\mathrm{P}_{\mathrm{ls}}^c)$ contains at most c connected components

The question remains on how to choose $\mathcal{S} \subset \mathcal{V}$. In view of our application in discrete tomography, one may use filtered back projection to receive a noisy reconstruction to manually choose $\mathcal{S}$. We illustrate this idea in Figure 5.5.

Remark 5.8. The models (P_0^c) and $(\mathrm{P}_{\mathrm{ls}}^c)$ are extendable so that the sinks are selected within branch and bound. However, there is an exponential number of optimal selections for the set of sinks $\mathcal{S}$ making this approach intractable.

In Section 5.1.2, we showed that the number of linear measurements for unique



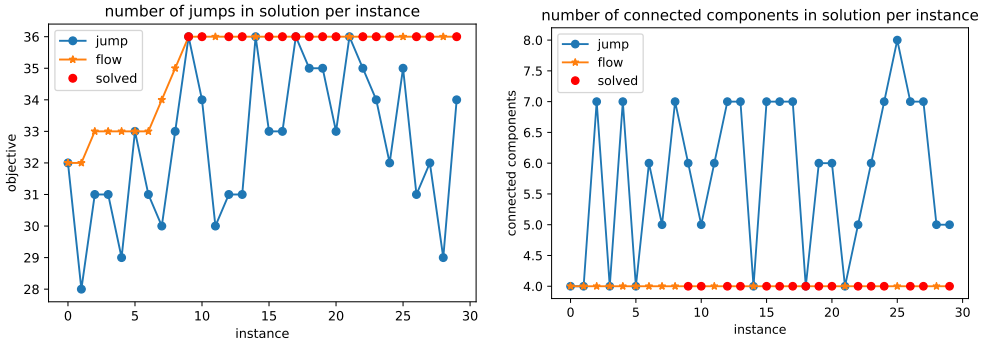
(a) filtered back projection (b) select sinks (c) ℓ_0 reconstruction

Figure 5.5: In (a) we used 180 angles and the filtered back projection to obtain a noisy but fast reconstruction. Then, we choose sinks (■) in (b), where each sink marks a potential connected component. Using (P_0^c) and only 10 angles we were able to find the correct solution (c) in less than a minute. Without predefining the sinks, we only found worse solutions within 10 minutes computing time.

reconstruction of an image depends either on the number of jumps or the connected components. To validate the theoretical findings from Section 5.1.2, we now devise a tractable experiment in order to carry out the optimization in (P_0^c). To this end, we constructed 30 images (9×9) with exactly 36 jumps and 4 connected components. Then, we reconstructed each image through (P_0^I) and (P_0^c) using two tomographic projections (horizontal and vertical direction). In view of (P_0^c), for each image we selected four sinks.

Figure 5.6 shows two plots analyzing the solutions obtained from the above mentioned reconstructions. As a result, we observe that for almost all instances (P_0^I) provided reconstructions with less jumps or more connected components implying that the target image was not reconstructed. On the other hand, (P_0^c) provided the optimal solution in many instances. Furthermore, (P_0^c) always provided solutions with more jumps or exactly 4 connected components.

We conclude that using connected components reduces the number of needed linear measurements also empirically. This coincides with the theory developed in Section 5.1.2.



(a) (b)

Figure 5.6: We created 30 images such that each image has exactly four connected components and 36 jumps. Then, we reconstructed each image (using two binary tomography projections) through (P_0^c) and (P_0^I), using GUROBI with timelimit of 300 seconds. Solutions marked (●) are only solved through (P_0^c). The plots above show that using flow constraints restricted the solution for each instance to four connected components. Consequently, most instances were uniquely defined by just two tomographic projections.

CHAPTER 6

Multilevel Optimization

Recall that tomographic projections correspond to line integrals, and a solution x of $Ax = b$ corresponds to a discrete solution, see Section 2.3. On account of this, varying the number of line integrals leads to different discretizations as the image discretization is in general adapted to the measurements resolution. Please see Figure 6.1 for an illustration. As a consequence, reconstructing an image from a finite number of line integrals corresponds to solving a linear system

$$A_h x = b_h, \tag{6.1}$$

where $h \in \mathbb{N}$ defines the resolution of the reconstructed image.

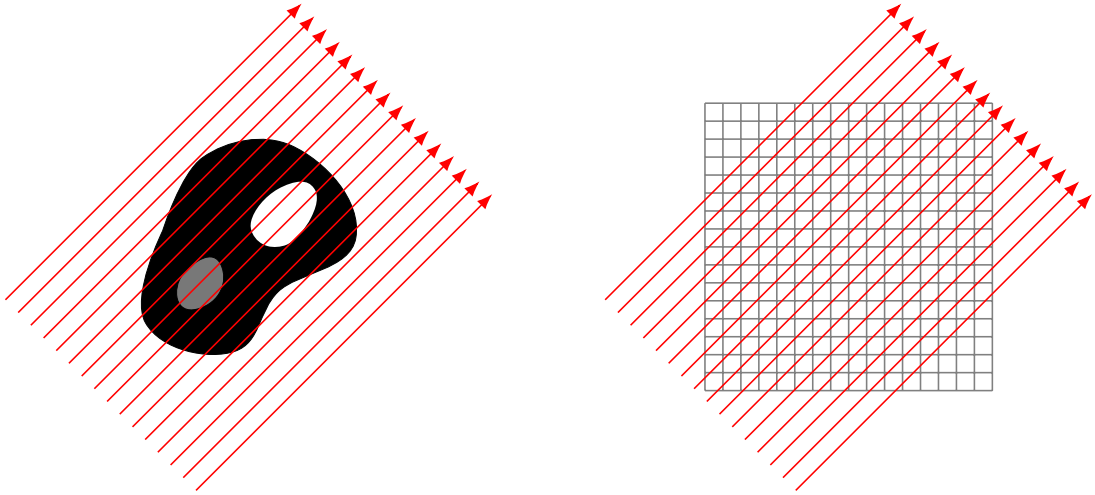
In contrast to the previous chapters, we here consider a different optimization model that incorporates the constraint (6.1) into the objective. It is given by

$$\inf_{x \in (0,1)^n} f_h(x) + \lambda g(x), \tag{6.2}$$

where f, g are a continuously differentiable convex functions such that $f_h(x) = 0$ if $A_h x = b_h$ and $f_h(x) > 0$ otherwise. Further, we require $\lambda > 0$. In particular, we consider the function $f_h(x) = \mathrm{KL}(A_h x, b_h)$[1].

In this chapter, we discuss how to use reconstructions with a coarse resolution to speed-up the reconstruction process for the desired resolution. In the context of differential equations, it is a well-known procedure to use different discretization levels to speed-up the convergence [BHM00]. Similarly, a multilevel approach to optimization problems was proposed in [Nas00]. Following [Nas00], multilevel optimization was further developed and applied to many other instances of optimization problems [GST08; Gra08; GT10; HPZ16; JH17; KM16; Nas00; Nas14; WG10].

[1] $\mathrm{KL}(x, y) = \sum_i x_i \log\left(\frac{x_i}{y_i}\right) + y_i - x_i$ for $x \geq 0$ and $y > 0$.



(a) 16 detectors (b) 16×16 grid

Figure 6.1: Illustration of the connection between the number of line integrals and the resulting discretization. (a) Line integrals in the continuous domain. In discrete tomography, we discretize the line integrals in (a) and obtain as many weighted sums over a grid equals the number of line integrals in (b).

In Section 6.1, we review the basics of the multilevel approach for unconstrained optimization problems. This results in the definition of a coarse correction term to calculate descent direction by coarse discretization level. Next, in Section 6.2, we discuss how multilevel optimization is applied to constrained optimization problems. We find that the correction term from the previous section remains unchanged, and only the constraints need to be handled separately. As the optimization problems considered in this chapter are defined on an open convex set, we can not apply classical line search methods, which minimize along gradient directions. For that reason, we introduce tools from *Differential Geometry* [Lee12; AMS08] in Section 6.3 which allows us to perform line search on manifolds. Finally, in Section 6.4, we illustrate the potential of the (proposed) multilevel approach by applying it to tomographic reconstructions.

6.1 Basics of Multilevel Optimization

We start with a general optimization problem in the calculus of variations[2]

$$\min_{u \in L^2(\Omega)} g(u), \tag{6.3}$$

where $\Omega \subset \mathbb{R}^2$ and $g : L^2(\Omega) \to \mathbb{R}$. Assuming that there are discretization levels, which approximate (6.3). We obtain a sequence of optimization problems given by

$$\min_{x \in \mathbb{R}^{n_\ell}} f_\ell(x), \tag{6.4}$$

where $f_\ell : \mathbb{R}^{n_\ell} \to \mathbb{R}$ is a smooth convex function. Let $\jmath < \ell$ denote two different discretizations, then $n_\jmath < n_\ell$. Therefore, f_0 is the coarsest discretization, e.g. $n_0 = 1$. For simplicity, we consider only two different discretizations $f : \mathbb{R}^n \to \mathbb{R}$ and $f_h : \mathbb{R}^{n_h} \to \mathbb{R}$ with $n > n_h$.

Multilevel optimization aims at accelerating iterative methods to solve (6.4) by employing coarser discretizations. The utilization of different discretization levels has already been used and is well-known in the context of differential equations [TSO01; BHM00]. Based on these results, Nash [Nas00] proposed a strategy to adopt this idea for smooth optimization.

An essential ingredient for multilevel optimization is the transition between fine and coarse grids, which we define as follows.

Definition 6.1. Let $P \in \mathbb{R}^{n \times n_h}$ and $R \in \mathbb{R}^{n_h \times n}$ such that

$$\sigma R^\top = P \tag{6.5}$$

for $\sigma > 0$. We call P the *prolongation map* and R the *restriction map*.

Generally, image patches, which not applied to every pixel, induce prolongation maps, and result in an image of lower resolution. As a result, a pixel on the coarse resolution is associated with an image patch on the fine resolution. This approach is illustrated in Figure 6.2.

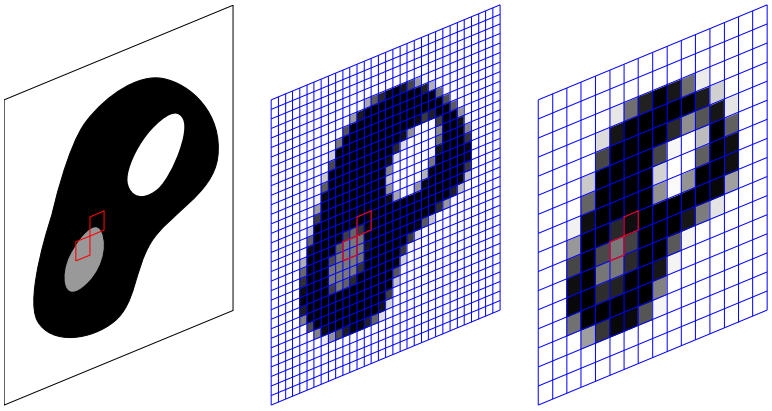
Next, we use a prolongation map in the sense of Definition 6.1 to transfer vectors from the coarse level to the fine level. Using this operator, we derive a new optimization problem at the current point x with fewer variables, i.e.,

$$\min_{h \in \mathbb{R}^{n_h}} f(x + Ph). \tag{6.6}$$

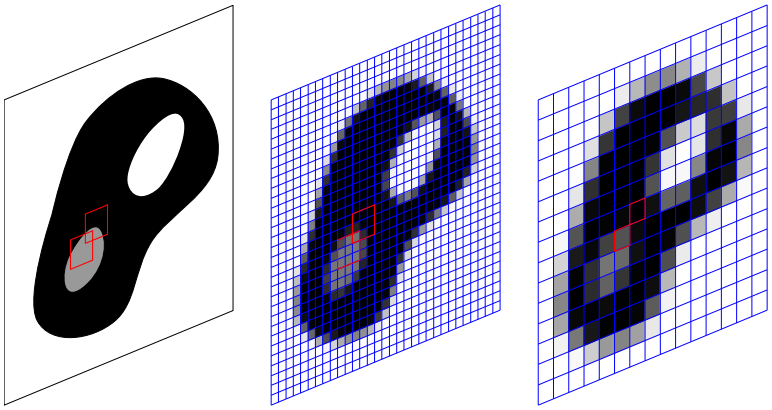
Problem (6.6) is related to *subspace minimization* [NW06]. However, in subspace minimization, one solves (6.6) approximately, which does not involve any information from coarser discretization. Hence, we consider the first Taylor expansion, which is a linear approximation of the function f at x concerning the subspace defined by P,

$$f(x + Ph) \approx f(x) + \langle \nabla f(x), Ph \rangle = f(x) + \langle R \nabla f(x), h \rangle. \tag{6.7}$$

[2] $L^2(\Omega)$ contains all functions $u : \Omega \to \mathbb{R}$ such that $\int_\Omega |u|^2 < \infty$



(a) non-overlapping patches



(b) overlapping patches

Figure 6.2: Both figures show discretizations of a continuous image using different techniques. Figure (a) averages over non-overlapping patches and Figure (b) uses a weighted averaging over overlapping patches to obtain discretizations. Considering the middle discretization from (a) or (b) as a fine grid, then each patch (red square) defines a pixel on the coarse grid (right).

To keep the formulas simple, we assume that $\sigma = 1$ in Definition 6.1. At this point, we use the coarse discretizations to obtain an error term for the first Taylor expansion which may depend on some coarse point

$$f_h(h; x_h) := f_h(x_h + h) - f_h(x_h) - \langle \nabla f_h(x_h), h \rangle. \tag{6.8}$$

Using (6.8), we construct a new function given by

$$\psi(h) := \langle R\nabla f(x), h \rangle + f_h(h; x_h). \tag{6.9}$$

Since f_h is convex and $f_h(x) \geq 0$ for all $x \in \mathbb{R}^{n_h}$, it follows immediately that $f_h(h; x_h)$

is convex for every $x_h \in \mathbb{R}^{n_h}$ and $f_h(h; x_h) \geq 0$ for all $h \in \mathbb{R}^{n_h}$.

Proposition 6.2. *Let ψ be defined by (6.9). Then it holds that*

$$\psi(h) < 0 \implies \langle Ph, \nabla f(x) \rangle < 0, \tag{6.10}$$

i.e., Ph is a descent direction for f at x.

Proof. Since $f_h(h; x_h) \geq 0$ for all $h \in \mathbb{R}^{n_h}$ and $\psi(h) < 0$ by assumption, it follows directly from construction that

$$0 > \psi(h) = f_h(h; x_h) + \langle R\nabla f(x), h \rangle.$$

Considering that $f_h(h; x_h) \geq 0$ it follows that

$$0 > \langle R\nabla f(x), h \rangle = \langle \nabla f(x), Ph \rangle,$$

which concludes the proof. □

Plugging (6.7) in (6.9) is called *first-order coherence* [GST08; WG10]. One obtains an n-th order coherence by using the n-th order Taylor expansion in (6.7) and use the n-th order error term in (6.9). However, in our cases, it is, in general, not tractable to explicitly calculate second-order derivatives.

In summary, multilevel optimization uses the surrogate model (6.9) with fewer variables to find a new descent direction. If (6.9) is bounded from below, we find a new descent direction by solving

$$h^* \in \operatorname*{arg\,min}_{h \in \mathbb{R}^{n_h}} \psi(h). \tag{6.11}$$

In the case of an unbounded ψ from (6.9), one does some gradient descent steps such that $\psi(h^*) < 0$. The resulting Ph^* provides in either way a descent direction for f at x.

Remark 6.1. The proof of Proposition 6.2 only requires that $R^\top = \sigma P$ with $\sigma > 0$. Therefore, one has many degrees of freedom to choose the prolongation map and restriction map, respectively.

The usage of a descent direction obtained by (6.11) is called *coarse correction*. In bulk of the publications [GST08; Gra08; GT10; HPZ16; JH17; KM16; Nas00; Nas14; WG10] the *coarse correction* is used to accelerate the convergence of existing descent algorithms.

Algorithm 6.1 summarize the ideas from [Nas00] adapted to our notation. In particular, the coarse correction step consists of finding a descent direction on a coarse level (line $2-4$) and the descent (line 5) along the obtained direction via line search. This procedure is used to accelerate the convergence of a descent algorithm by applying it between iterations.

In Figure 6.3, we illustrate a simple example of the power of multilevel acceleration. The structure of this simple example also fits a lower resolution. As a consequence,

Algorithm 6.1: Coarse Correction Step

Input: f fine level function, f_h coarse function,
x current point, x_h coarse point
Output: $\overline{x}$ coarse correction with $f(\overline{x}) < f(x)$

Function `CoarseCorrection`(f, f_h, x, x_h)
  $\psi(h) := \langle R\nabla f(x), h\rangle + f_h(h; x_h)$ `/* define coarse function */`
  find h^* such that $\psi(h^*) < 0$ `/* e.g. gradient descent */`
  $d \leftarrow Ph^*$ `/* descent direction */`
  $\overline{x} \leftarrow x + \alpha d$ `/*` $\alpha > 0$ `such that` $f(x + \alpha d) < f(x)$ `*/`
  return $\overline{x}$
end

multilevel propagates this coarse structure to the target resolution. Comparing the multilevel iterations to the gradient descent, we observe that the gradient descent struggles to reconstruct the coarse structure compared to the multilevel approach.

In practice, it is difficult to predict precisely whether the coarse correction provides a lower objective value than direct gradient descent at the fine level. For instance, if $R\nabla f(x)$ is componentwise approximately zero, the coarse model might give only a small descent applied to the fine level. Thus, [GST08] proposed a condition for a coarse correction step:

$$\|R\nabla f(x)\| \geq \kappa \|\nabla f(x)\| \text{ and } \|R\nabla f(x)\| > \varepsilon, \tag{6.12}$$

where $\kappa \in (0, \min\{1, \|R\|\})$ and $\varepsilon > 0$. By condition (6.12) a coarse correction is rejected whenever the coarse gradient $R\nabla f(x)$ is small or small compared to the fine gradient $\nabla f(x)$.

6.2 Constrained Multilevel Optimization

So far, we have discussed multilevel optimization in the unconstrained optimization setting. However, in previous chapters, we have discussed that the inclusion of bounds leads to a significant reduction in the number of measurements for exact recovery. This motivates a multilevel approach that is also able to handle variable bounds.

We proceed as follows. First, we examine a general smooth convex optimization problem defined on different levels of discretization. Based on the idea of Nash [Nas10; Nas14], we design a coarse correction model for two instances, similar to the case of the unconstrained optimization. Continuing with Nash's approach, it becomes apparent that we require different restriction maps for the gradient and the constraints. Then, we specialize in variable bounds and review the result from [Gra08] providing a suitable restriction map.

We start with the definition of a general smooth convex constrained optimization

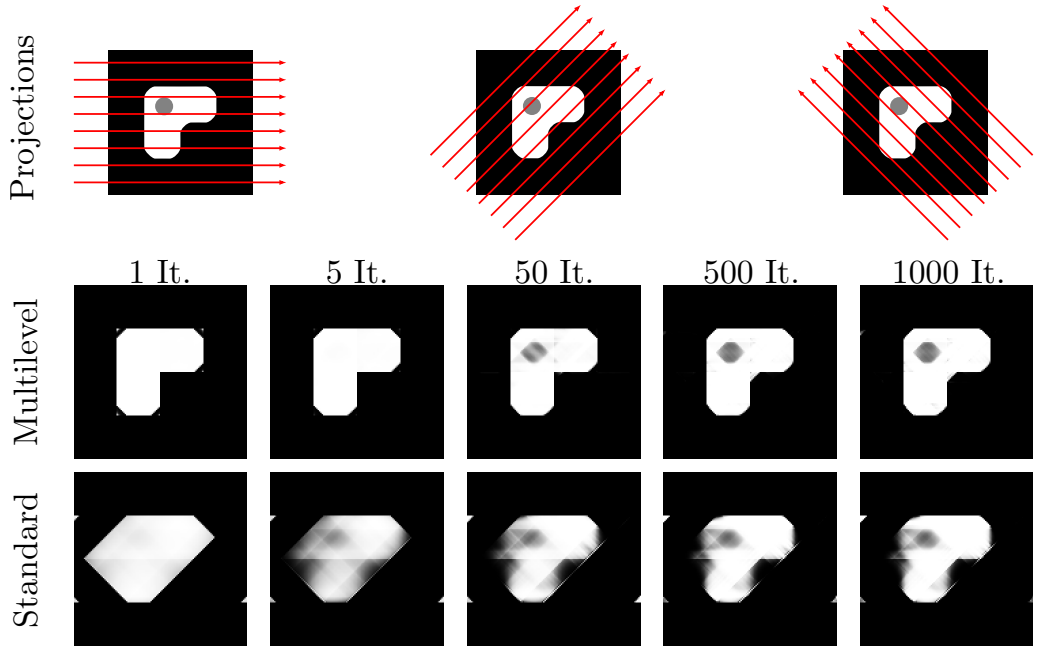


Figure 6.3: Illustration of the coarse correction on a low complexity image using only three projections. The first row illustrates the projections on the original image. We used 128 rays for each projection and measured the image on the resolution 4096×4096. Afterwards, we reconstructed the image on a 128×128 grid via $\min_{x \in (0,1)^n} \mathrm{KL}(Ax, b)$ together with gradient descent. The second row shows the result of using coarse correction steps on a 4×4 grid during the iterations of gradient descent, where the last row is the reference without coarse correction. As a result, we see that the coarse correction step shifts the fine iterations to a solution fitting to the measurements on the coarse grid.

problem

$$\begin{aligned} &\min\ f(x) \\ &\text{s.t. } p(x) = 0, \\ &\quad\ q(x) \leq 0, \end{aligned} \tag{6.13}$$

where $f : \mathbb{R}^n \to \mathbb{R}, p : \mathbb{R}^n \to \mathbb{R}^{m_p}, q : \mathbb{R}^n \to \mathbb{R}^{m_q}$ are smooth convex functions so that the constraints define a closed convex set. As in the unconstrained case, there exists a related optimization problem

$$\begin{aligned} &\min_x\ f_h(x) \\ &\text{s.t. } p_h(x) = 0, \\ &\quad\ q_h(x) \leq 0 \end{aligned} \tag{6.14}$$

on a coarser discretization level which approximates the problem (6.13). The functions f_h, p_h, q_h are likewise smooth, convex and the constraints define a closed convex set.

Following Nash [Nas10; Nas14], we use the Lagrange function to design an analog coarse correction model, as in the last section. For simplicity, we first omit the inequality constraints in (6.13) and (6.14), .i.e., we consider

$$\begin{aligned} &\min_x \ f(x) \\ &\text{s.t. } \ p(x) = 0, \end{aligned} \tag{6.15}$$

$$\begin{aligned} &\min_x \ f_h(x) \\ &\text{s.t. } \ p_h(x) = 0. \end{aligned} \tag{6.16}$$

Let $\mathcal{L}(x, \lambda)$ be the Lagrange function of (6.15) and $\mathcal{L}^h(x^h, \lambda^h)$ the Lagrange function of (6.16). Further, we define by $\mathcal{L}^h(h; x^h, \lambda^h)$ the adapted error term as in (6.8). As a result, we obtain an analogous coarse correction term given by

$$\psi(h, \tilde{h}) := \left\langle \begin{pmatrix} R^\nabla & 0 \\ 0 & R^= \end{pmatrix} \nabla \mathcal{L}(x, \lambda), \begin{pmatrix} h \\ \tilde{h} \end{pmatrix} \right\rangle + \mathcal{L}^h \left(\begin{pmatrix} h \\ \tilde{h} \end{pmatrix}; x^h, \lambda^h \right). \tag{6.17}$$

as in the unconstrained case.

We assume that $p(x) = 0$, $p_h(x_h) = 0$, i.e., x is feasible for (6.15) and x_h is feasible for (6.16), and set $\lambda = \lambda^h = 0$. The minimization of (6.17) w.r.t. $h, \tilde{h}$ by means of the optimality condition $\nabla \psi(h, \tilde{h}) = 0$ implies the convex optimization problem

$$\begin{aligned} &\min_h \ \langle R^\nabla \nabla f(x), h \rangle + f_h(h; x^h), \\ &\text{s.t. } \ p_h(x^h + h) = 0, \end{aligned} \tag{6.18}$$

which computes a coarse correction direction for problem (6.15). In particular, a solution of problem (6.18) yields a descent direction $P^\nabla h^*$ for f at x and h^* is a feasible direction in the solution space of (6.18). The operator $P^\nabla h^*$ is the related prolongation map to R^∇. If $P^\nabla h^*$ is not a feasible direction in the solution space of (6.15), we need to apply an orthogonal projection on the solution space.

By introducing slack variables in (6.13) and (6.14), we perform an analogous calculation as above and obtain a coarse correction model including inequality constraints

$$\begin{aligned} &\min \ \langle R^\nabla \nabla f(x), h \rangle + f_h(h; x^h) \\ &\text{s.t. } \ p_h(x^h + h) = 0, \\ &\quad q_h(x^h + h) - q_h(x^h) \leq -R^\leq q(x). \end{aligned} \tag{6.19}$$

It follows from (6.19) that by including inequality constraints, we need an additional restriction map besides R^∇.

With (6.19) we completed the design of the coarse correction model for constrained optimization. Next, we specialize in variable bounds meaning that $x \in [l, u]$ with

$l, u \in \mathbb{R}^n$ and $l < u$, resp. $x \in [l_h, u_h]$ with $l, u \in \mathbb{R}^{n_h}$ and $l_h < u_h$. That means we have two functions $q^l(x) = l - x$, $q^u(x) = x - u$ resp. $q_h^l(x) = l_h - x$, $q_h^u(x) = x - u_h$ realizing the inequality constraints. Utilizing (6.19), we obtain for variable bounds the following coarse correction model

$$\begin{aligned} &\min_h \ \langle R^\nabla \nabla f(x), h\rangle + f_h(h; x^h) \\ &\text{s.t. } R_l^{\leq}(l - x) \leq h \leq R_u^{\leq}(u - x). \end{aligned} \tag{6.20}$$

As mentioned above, a solution h^* induces a descent direction by $P^\nabla h$ for f at x, but it may be infeasible for the solution space in the fine level meaning

$$x + \alpha P^\nabla h \notin [l, u], \qquad \forall \alpha > 0. \tag{6.21}$$

We use restriction maps as proposed by [Gra08], defined below, to avoid (6.21). Adapting the restriction maps from [Gra08] to our notation, we gain the following component-wise definition

$$(l_h)_j = \frac{1}{\|P^\nabla\|_\infty} \max_{i=1,\dots,n} \begin{cases} (l-x)_i & P_{ij}^\nabla > 0 \\ (x-u)_i & P_{ij}^\nabla < 0 \end{cases}, \tag{6.22}$$

$$(u_h)_j = \frac{1}{\|P^\nabla\|_\infty} \min_{i=1,\dots,n} \begin{cases} (u-x)_i & P_{ij}^\nabla > 0 \\ (x-l)_i & P_{ij}^\nabla < 0 \end{cases}, \tag{6.23}$$

where P^∇ is the corresponding prolongation map to the restriction map R^∇ as in Definition 6.1, $l_h = R_l^{\leq}(l - x)$ and $u_h = R_u^{\leq}(x - u)$.

Lemma 6.3 ([Gra08, Lem. 4.3])**.** *Consider the variable bounds $l, u \in \mathbb{R}^n$ with $l < u$ and $x \in \mathbb{R}^n$ at the fine level. Let l_h and u_h be defined by* (6.22)*. Then, it holds*

$$l \leq x + P^\nabla h \leq u, \qquad \forall h : l_h \leq h \leq u_h. \tag{6.24}$$

Proof. Define $\phi_i = \sum_j |P_{ij}^\nabla|$ as the sum of the absolute values of the i-th row of P^∇ and calculate

$$x_i + \sum_j P_{ij}^\nabla h_j \tag{6.25}$$

$$= x_i + \sum_{j, P_{ij}^\nabla < 0} |P_{ij}^\nabla|(-h_j) + \sum_{j, P_{ij}^\nabla > 0} |P_{ij}^\nabla| h_j \tag{6.26}$$

$$\geq x_i + \sum_{j, P_{ij}^\nabla < 0} |P_{ij}^\nabla| \frac{-\min_i (x-l)_i}{\|P^\nabla\|_\infty} + \sum_{j, P_{ij}^\nabla > 0} |P_{ij}^\nabla| \frac{\max_i (l-x)_i}{\|P^\nabla\|_\infty} \tag{6.27}$$

$$\geq x_i + \sum_{j, P^{\nabla}_{ij} < 0} |P^{\nabla}_{ij}| \frac{(l-x)_i}{\|P^{\nabla}\|_\infty} + \sum_{j, P^{\nabla}_{ij} > 0} |P^{\nabla}_{ij}| \frac{(l-x)_i}{\|P^{\nabla}\|_\infty} \tag{6.28}$$

$$= x_i + \phi_i \frac{(l-x)_i}{\|P^{\nabla}\|_\infty} \tag{6.29}$$

$$= \frac{\phi_i}{\|P^{\nabla}\|_\infty} l_i + \left(1 - \frac{\phi_i}{\|P^{\nabla}\|_\infty}\right) x_i \geq l_i. \tag{6.30}$$

The last inequality follows from $x_i \geq l_i$ and $\frac{\phi_i}{\|P^{\nabla}\|_\infty} \in [0,1]$. Analogously, we obtain that

$$x_i + \sum_j P^{\nabla}_{ij} h_j \leq u_i. \tag{6.31}$$

□

Remark 6.2. The authors Kočvara and Mohammed [KM16] have already successfully adopted the restriction map from [Gra08] to optimize quadratic and non-quadratic obstacle problems. Their results motivated our analysis and approach in this chapter.

Finally, we review the two primary results of this section. We constructed a coarse correction model (6.20) analogous to the unconstrained case with variable bounds. This enables us to compute a descent direction $P^{\nabla} h^*$ of f at x by optimizing on a coarse discretization level. We have remarked that without a special choice of the prolongation map $R^{\leq}$, it may occur that $P^{\nabla} h^*$ is infeasible for the solution space at the fine level, see (6.21). Choosing the prolongation map according to (6.22), we ensure that

$$x + \alpha P^{\nabla} h \in [l, u], \qquad \forall \alpha \in [0,1], \tag{6.32}$$

which is proved by Lemma 6.3. Consequently, unless the solution $h^* = 0$ for (6.20), we are able to achieve a descent of f along $P^{\nabla} h^*$ at x.

6.3 Optimization on Manifolds

The optimization problems in this chapter consist of minimizing a smooth convex function over an open convex set. Thus, classical line search methods in the Euclidean space cannot be applied. As a remedy, we regard the open convex set as a manifold which provides tools to perform a modified variant of classical line search methods.

In the preliminaries in Section 2.1.2, we briefly introduced iterative gradient descent on manifolds. Using this procedure, we need two ingredients:

(i) a Riemannian structure on the manifold and

(ii) a retraction which maps tangent vectors to the manifold.

Our primary goal in this section is to equip the open box manifold $\mathcal{M} = (l, u)$ with (i) and (ii), where $l, u \in \mathbb{R}^n$ and $l < u$. Hence, in Section 6.3.1, we apply the result from [ABB04] to achieve (i), which provides a construction to define a Riemannian structure on open convex sets. The construction of a retraction is in general difficult. Therefore, we introduce in Section 6.3.2 a concept to design a retraction through an existing retraction from another manifold.

6.3.1 Open Convex Sets as Riemannian Manifolds

In the following, we furnish the box manifold $\mathcal{M} = (l, u)$ with a Riemannian structure employing the construction of [ABB04]. We remark that any manifold[3] admits a Riemannian structure [AMS08].

For a start, we bring to mind the crucial terms from the preliminaries in Section 2.1.2. Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function defined on a manifold $\mathcal{M}$. Then, the differential $\mathrm{d}_x f : T_x\mathcal{M} \to \mathbb{R}$ generalizes the directional derivative along a direction defined w.r.t. the tangent space $T_x\mathcal{M}$. We showed in the preliminaries that if $\mathcal{M}$ is a n-dimensional manifold, each tangent space is isomorphic to $\mathbb{R}^n$.

Further, let $\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ be a Riemannian metric, which is an inner product on the tangent space. Then, it exists a unique element $\nabla_{\mathcal{M}} f(x) \in T_x\mathcal{M}$ called the Riemannian gradient such that ([AMS08, p. 46])

$$\mathrm{d}_x f[\xi] = \langle \nabla_{\mathcal{M}} f(x), \xi \rangle_x, \qquad \forall \xi \in T_x\mathcal{M}. \tag{6.33}$$

In view of our setting, we are able to calculate the Euclidean gradient $\nabla f(x)$ and by the identification $T_x\mathcal{M} \cong \mathbb{R}^n$, it follows that

$$\langle \nabla_{\mathcal{M}} f(x), h \rangle_x = \langle \nabla f(x), h \rangle, \qquad \forall h \in \mathbb{R}^n\,. \tag{6.34}$$

As a consequence, the coarse correction term (6.9) is still valid for calculating descent directions on a coarse level, even if we optimize on a manifold.

Having reviewed the essential terminology, we construct a Riemannian metric for the box manifold (l, u) using the instructions from [ABB04]. To this end, we need to define a function $H : \mathcal{M} \to \mathbb{S}^n_{++}$, where $\mathbb{S}^n_{++}$ is set of positive definite $n \times n$ matrices, such that

$$\langle u, v \rangle_x = \langle u, H(x)v \rangle, \qquad u, v \in T_x\mathcal{M} \subset \mathbb{R}^n \tag{6.35}$$

$$\nabla_{\mathcal{M}} f(x) = H(x)^{-1} \nabla f(x), \qquad \nabla f(x) \in T_x\mathcal{M}. \tag{6.36}$$

The function H is defined as the Hessian of a Legendre type function.

Definition 6.4 ([Roc97, Chapter 26])**.** A lower semicontinuous, proper and convex function $f : C \subset \mathbb{R}^n \to \mathbb{R}^n \cup \{\infty\}$ defined on a closed convex set C is called

(i) essentially smooth if h is differentiable on $\mathrm{int}(\mathrm{dom}(f))$ and $|\nabla f(x^j)| \to \infty$ for any sequence x^j converging to the boundary.

[3] We emphasize that as defined in the preliminaries, we only consider smooth manifolds but only write manifold.

(ii) of Legendre type if f is essentially smooth and strictly convex on $\operatorname{int}(\operatorname{dom}(f))$.

Designing a Riemannian structure by the aid of (6.35) and (6.36), we define the following Legendre type function

$$h(x) = \langle u - l, (x - l)\log(x - l) + (u - x)\log(u - x)\rangle. \tag{6.37}$$

The Hessian of h is given by

$$H(x) := \nabla^2 h(x) = \operatorname{diag}\left(\frac{(x - l)(u - x)}{(u - l)^2}\right) \tag{6.38}$$

Next, we insert (6.38) in (6.35) and (6.36). This yields the Riemannian metric

$$\langle u, v\rangle_x := \left\langle u, \frac{(x - l)(u - x)}{(u - l)^2} v \right\rangle, \qquad u, v \in T_x(l, u), \tag{6.39}$$

and the related Riemannian gradient

$$\nabla_{\mathcal{M}} f(x) = \frac{(u - l)^2}{(x - l)(u - x)} \nabla f(x). \tag{6.40}$$

The choice of (6.37) seems arbitrary. We will later see that when defining the retraction in the following section, we notice the advantage of precisely has chosen this function.

6.3.2 Retraction for Box Constraints

In this section, we construct a retraction $\mathfrak{R}_x : T_x\mathcal{M} \to \mathcal{M}$, which generalizes the notion of moving along a line on manifolds. Regarding our case, we let $x \in (l, u)$ and $0 \neq v \in \mathbb{R}^n$. Then, if we follow the line $x + \alpha \cdot v$ it eventually occurs that $x + \alpha \cdot v \notin (l, u)$ for a sufficiently large $\alpha > 0$. Using a retraction, we always stay on the manifold (l, u) meaning $\mathfrak{R}_x(\alpha \cdot v) \in (l, u)$ for all $\alpha \geq 0$.

For the definition of a retraction, we refer to Section 2.1.2. Generally, an approach to retrieve a retraction is to define an additional differential structure called *affine connection* [AMS08, Sec. 5.2]. This is used to define curves with zero acceleration corresponding to straight lines in the Euclidean space. Such curves are called *geodesics* [AMS08, Sec. 5.4]. Furthermore, geodesics are used to define the so-called *exponential map* [AMS08, Sec. 5.4], which is also a retraction if it is defined on the whole tangent space [AMS08, Prop. 5.4.1].

Instead of constructing a retraction from scratch, we show how to reuse an existing retraction by transferring it to the desired manifold. Later in this section, we consider the following retraction from [GHW20, Sec. 8.2.1.2]

$$\mathfrak{R}_x^{\Delta_2}(\xi) = \frac{x e^{\frac{\xi}{x}}}{\langle x, e^{\frac{\xi}{x}}\rangle}, \qquad \xi \in T_x\Delta_2, \tag{6.41}$$

which is defined on the simplex manifold Δ_n[4]. Indeed, this retraction is the exponential map induced by the e-connection [Ay17, Prop. 2.5].

Next, we proceed with the main result of this section. For that reason, let $\mathfrak{R}^{\mathcal{N}}$ be a known retraction on the manifold $\mathcal{N}$. We are interested in constructing a retraction $\mathfrak{R}^{\mathcal{M}}$ for the manifold $\mathcal{M}$. Suppose, we have a diffeomorphism $F : \mathcal{M} \to \mathcal{N}$ between these manifolds. Accordingly, we define $\mathfrak{R}^{\mathcal{M}}$ as in the following diagram.

$$\begin{array}{ccc} T_x\mathcal{M} & \xrightarrow{\mathfrak{R}^{\mathcal{M}}_x} & \mathcal{M} \\ {\scriptstyle \mathrm{d}_x F}\downarrow & & \uparrow{\scriptstyle F^{-1}} \\ T_{F(x)}\mathcal{N} & \xrightarrow[\mathfrak{R}^{\mathcal{N}}_{F(x)}]{} & \mathcal{N} \end{array} . \tag{6.42}$$

Within the next proposition, we prove that (6.42) indeed commutes.

Proposition 6.5. *Let $F : \mathcal{M} \to \mathcal{N}$ be a diffeomorphism between two manifolds. Having a retraction $R_x : T_x\mathcal{N} \to \mathcal{N}$, the map $\hat{R}_x : T_x\mathcal{M} \to \mathcal{M}$ defined by*

$$\hat{R}_x(v) = F^{-1}(R_{f(x)}(\mathrm{d}_x F[v])) \tag{6.43}$$

is also a retraction.

Proof. We need to check, that $\hat{R}_x$ is smooth and that

- $\hat{R}_x(0) = x$ and
- $\mathrm{d}_0\hat{R}_x = \mathrm{Id}$.

$\hat{R}_x$ is smooth because it is a composition of smooth functions. By assumption, F is a diffeomorphism. Then $\mathrm{d}_x F$ is a linear isomorphism [AMS08, Prop. 3.6] and consequently

$$\mathrm{d}_x F[0] = 0. \tag{6.44}$$

Therefore, we calculate

$$\begin{aligned} \hat{R}_x(0) =& F^{-1}(R_{F(x)}(\mathrm{d}_x F[0])) && (6.45)\\ =& F^{-1}(R_{F(x)}(0)) && (6.46)\\ =& F^{-1}(F(x)) = x. && (6.47) \end{aligned}$$

Let $A : V \to W$ be a linear map between two vector spaces V, W. The differential at $v \in V$ in direction $\eta \in V$ is given by

$$\mathrm{d}_v A[\eta] = \frac{\mathrm{d}}{\mathrm{d}t} A(v + t \cdot \eta)\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} Av + t \cdot A\eta\Big|_{t=0} = A\eta. \tag{6.48}$$

As a consequence, we obtain $\mathrm{d}_v A = A$ and consequently $\mathrm{d}_0(\mathrm{d}_x F) = \mathrm{d}_x F$. This result

[4] $\Delta_n = \{x \in \mathbb{R}^n \mid \langle \mathbb{1}, x\rangle = 1, x > 0\}$

is needed to show the second property of the retraction:

$$\begin{aligned} \mathrm{d}_0\hat{R}_x &= \mathrm{d}_{F(x)}F^{-1} \circ \mathrm{d}_{\mathrm{d}_x F[0]}R_{f(x)} \circ \mathrm{d}_0(\mathrm{d}_x F) && (6.49)\\ &= (\mathrm{d}_x F)^{-1} \circ \mathrm{d}_0 R_{F(x)} \circ \mathrm{d}_x F && (6.50)\\ &= (\mathrm{d}_x F)^{-1} \circ \mathrm{Id} \circ \mathrm{d}_x F && (6.51)\\ &= \mathrm{Id}. && (6.52) \end{aligned}$$

□

Employing the result of Proposition 6.5, we may transfer the retraction from (6.41) to the box manifold (l, u). But first, we need to define a diffeomorphism $F : (l, u) \to \nabla_2$,

$$\begin{aligned} F : (l,u) &\to \Delta_2, \\ x &\mapsto \frac{1}{u-l}\begin{pmatrix} u-x \\ x-l \end{pmatrix}, \\ F^{-1} : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &\mapsto x_1 \cdot l + (1-x_1)\cdot u. \end{aligned} \tag{6.53}$$

As F, F^{-1} are affine linear functions, they are smooth. Moreover, $(F^{-1} \circ F)(x) = x$ and $(F \circ F^{-1})(x) = x$. Hence, F is a diffeomorphism between (l, u) and Δ_2. We also need to calculate the differential $\mathrm{d}_x F$. This is done by defining a smooth curve $\gamma : \mathbb{R} \to (l, u)$ with $\gamma(0) = x$ and $\dot{\gamma}(0) = \xi \in T_x(l, u)$

$$\begin{aligned} \mathrm{d}_x F[\xi] = \frac{\mathrm{d}}{\mathrm{d}t}(F \circ \gamma)(t)\Big|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{u-l}\begin{pmatrix} u-\gamma(t) \\ \gamma(t)-l \end{pmatrix}\Big|_{t=0} && (6.54)\\ &= \frac{1}{u-l}\begin{pmatrix} -\dot{\gamma}(0) \\ \dot{\gamma}(0) \end{pmatrix} && (6.55)\\ &= \frac{1}{u-l}\begin{pmatrix} -\xi \\ \xi \end{pmatrix}. && (6.56) \end{aligned}$$

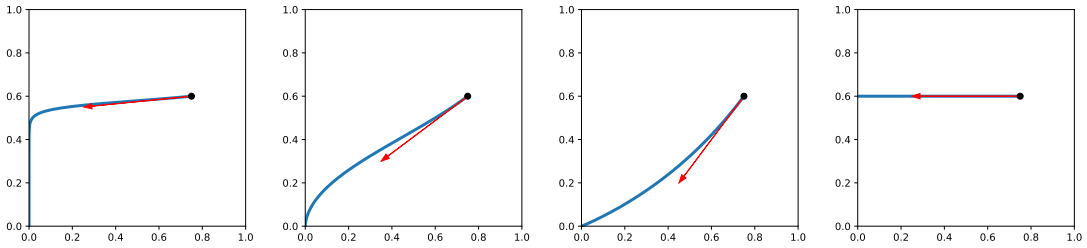
As $\mathrm{d}_x F$ is a linear isomorphism, we deduce that $T_{f(x)}\Delta_2 = \{v \in \mathbb{R}^2 \ : \ \langle \mathbb{1}, v\rangle = 0\}$.

Using Proposition 6.5 and F defined as above, we construct $\mathfrak{R}_x^{(l,u)}$ by

$$\mathfrak{R}_x^{(l,u)}(v) := f^{-1}(\mathfrak{R}_{f(x)}^{\Delta_2}(\mathrm{d}f(x)[v])) = \frac{(u-l)(x-l)e^{\frac{(u-l)^2 v}{(x-l)(u-x)}}}{u-x+(x-l)e^{\frac{(u-l)^2 v}{(x-l)(u-x)}}} + l. \tag{6.57}$$

All operations in the above equation are understood component-wise.

We mentioned in the last section, that our particular choice of a Riemannian structure is less arbitrary than it seems. To this end, let $f : (0, 1)^n \to \mathbb{R}$ be a smooth convex function and $\nabla f(x)$ the corresponding Euclidean gradient at $x \in (0, 1)^n$. The



(a) (b) (c) (d)

Figure 6.4: Illustration of the behavior of the retraction (6.57) for different directions in the case of $\mathcal{M} = (0,1)^2$. The red arrows are the tangent vectors identified with $\mathbb{R}^2$ and the blue line shows the computation of the retraction in direction of the tangent vector. Unless the tangent vector is orthogonal to a face of the box (see (d)), the curve induced by the retraction converges to a corner of the box.

Riemannian gradient according to our choice is given by

$$\nabla_{(0,1)^n} f(x) = \frac{\mathbb{1}}{x \cdot (\mathbb{1} - x)} \nabla f(x). \tag{6.58}$$

Then, a line search in direction of the steepest descent meaning $-\nabla f(x)$ for some step size $\alpha > 0$ results in

$$\mathfrak{R}_x^{(0,1)}(-\alpha \nabla f(x)) = \frac{x \cdot e^{-\alpha \nabla f(x)}}{\mathbb{1} - x + x \cdot e^{-\alpha \nabla f(x)}}. \tag{6.59}$$

We recognize that the factor of the Riemannian gradient and the retraction cancel out, which makes the retraction easier to compute.

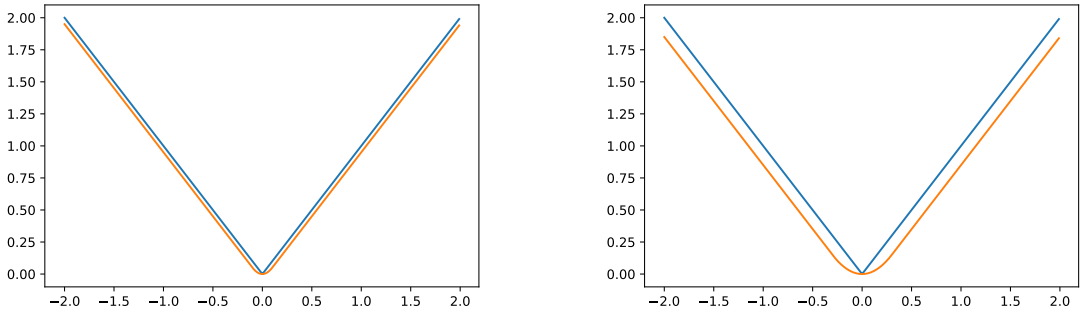
In Figure 6.4, we illustrate the behavior of the retraction (6.57) for different directions in the case of $\mathcal{M} = (0,1)^2$.

6.4 Experiments

In this section, we illustrate the potential of our multilevel framework by applying it to discrete tomography. To this end, we consider the following optimization problem

$$\inf_{x \in (0,1)^n} \underbrace{\mathrm{KL}(Ax, b) + \lambda \langle g_\rho(\nabla x), \mathbb{1} \rangle}_{=: f(x)}, \tag{6.60}$$

where g_ρ is the componentwise smooth approximation of the absolute value function and is given by



(a) $\rho = 0.1$ (b) $\rho = 1$

Figure 6.5: The plots show the Huber loss (6.61) (orange) and the absolute value function (blue). This illustrates that for decreasing values of ρ the Huber loss smoothly approximates the absolute value function.

$$g_\rho(x)_i = \begin{cases} \frac{x_i^2}{2\rho} & \text{if } |x| < \rho, \\ |x_i| - \frac{\rho}{2} & \text{otherwise} \end{cases}. \tag{6.61}$$

We note that g_ρ is continuously differentiable and is known as the *Huber Loss function* [Hub64]. Thus, the term $\langle g_\rho(\nabla x), \mathbb{1} \rangle$ is a smooth approximation of $\|\nabla x\|_1$. In Figure 6.5, we compare the absolute value function and the Huber loss function for $n = 1$. We observe that for decreasing values of ρ the Huber loss smoothly approximate the absolute value function.

We recall that when ∇ is applied to a vector x, it means the two-dimensional discrete gradient, and when ∇ is applied to $f(x)$, it means the gradient of f at x.

We approach (6.60) with gradient descent on manifolds as presented in Section 2.1.2. Since the considered objective is smooth, convex and $(0,1)^n$ is an open convex set, the authors in [ABB04] proved that gradient descent converges to the minimal objective value.

As described in the introduction of this chapter, we assume that the optimization problem (6.60) exists on different discretization levels indexed by h, i.e.,

$$\inf_{x \in (0,1)^{n_h}} \underbrace{\mathrm{KL}(A_h x, b_h) + \lambda_h \langle g_\rho(\nabla x), \mathbb{1} \rangle}_{=: f_h(x)}. \tag{6.62}$$

Before we adapt the coarse correction term

$$\psi(d) = \langle R \nabla f(x), d \rangle + f_h(d; x_h) \tag{6.63}$$

to our particular case, we prove the following preparatory lemma.

Lemma 6.6. *Let $f(x) = \mathrm{KL}(Ax, b)$ with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Then it holds*

$$f(h; x) = f(x + h) - f(x) - \langle \nabla f(x), h\rangle = \mathrm{KL}(A(x + h), Ax). \tag{6.64}$$

Proof. The following straightforward computation shows the assertion of the lemma:

$$\begin{aligned}
& f(x+h) - f(x) - \langle \nabla f(x) \\
=\ & \mathrm{KL}(A(x+h), b) - \mathrm{KL}(Ax, b) - \left\langle A^T \log\left(\frac{Ax}{b}\right), h\right\rangle \\
=\ & \left\langle A(x+h), \log\left(\frac{A(x+h)}{b}\right)\right\rangle - \left\langle Ax, \log\left(\frac{Ax}{b}\right)\right\rangle - \langle \mathbb{1}, Ah\rangle - \left\langle \log\left(\frac{Ax}{b}\right), Ah\right\rangle \\
=\ & \left\langle Ax, \log\left(\frac{A(x+h)}{Ax}\right)\right\rangle + \left\langle Ah, \log\left(\frac{A(x+h)}{Ax}\right)\right\rangle - \langle \mathbb{1}, Ah\rangle \\
=\ & \left\langle A(x+h), \log\left(\frac{A(x+h)}{Ax}\right)\right\rangle + \langle \mathbb{1}, Ax - A(x+h)\rangle \\
=\ & \mathrm{KL}(A(x+h), Ax).
\end{aligned}$$

□

By Lemma 6.6, the coarse correction term is given by

$$\psi_h(d) = \langle R^\nabla \nabla f(x), d\rangle + \mathrm{KL}(A_h(x_h + d), A_h x_h) + \lambda_h \langle \mathbb{1}, g_{\rho,h}(\nabla d; x_h)\rangle. \tag{6.65}$$

The crucial implication of Lemma 6.6 is that b_h does not appear in the coarse correction term. As a consequence, when applying the multilevel approach to tomographic reconstruction, we only need the projection data at the finest level.

The coarse correction term is not defined for all $d \in \mathbb{R}^{n_h}$. Hence, we use the restriction map defined in (6.22) to ensure that $x_h + d \in (0, 1)^n$. As a result, we gain the following coarse correction model

$$\inf_{d\in(l_h,u_h)} \psi_h(d). \tag{6.66}$$

If we find $d^* \neq 0$ by the above model so that $\psi_h(d^*) < 0$, then Proposition 6.2 ensures that $P^\nabla d^*$ is a descent direction of f at x. Moreover, using the restriction map defined in (6.22) yields

$$x + P^\nabla d^* \in (0, 1)^n. \tag{6.67}$$

To summarize, we compute a coarse correction step (6.67) in between gradient descent steps on the manifold. This defines a gradient descent algorithm, which uses two levels, as described in Algorithm 6.2. In the gradient descent part, we apply Algorithm 2.1, which uses the retraction (6.57) to obtain a descent direction on the manifold. Using Algorithm 6.2 inside the coarse correction step of Algorithm 6.1 to find a new coarse correction, see line 3, we achieve a *Multilevel Gradient Descent*.

In the following numerical experiment, we used the *full weighting* [TSO01] prolon-

Algorithm 6.2: Two Level Optimization

Input: fine level function f, coarse function f_h, starting point x
Output: x^*

Function `Multilevel`(x, f, f_h)
  $\overline{x} \longleftarrow x$
  while $\overline{x}$ *is not optimal* **do**
    $\overline{x}_h \longleftarrow R\overline{x}$
    if $\|R^\nabla \nabla f_h(\overline{x})\| \geq \kappa \|\nabla f(\overline{x}_h)\|$ *and* $\|R^\nabla \nabla f_h(\overline{x}_h)\| \geq \varepsilon$ **then**
      $\overline{x} \longleftarrow$ `CoarseCorrection`$(f, f_h, \overline{x}, \overline{x}_h)$
    end
    $\overline{x} \longleftarrow$ `GradientDescent`$(f, \overline{x})$
  end
  return $\overline{x}$
end

gation map P which is illustrated in Figure 6.6. Specifically, we defined $x_h := \frac{1}{16} P^\top x$, $P^\nabla := \frac{1}{4} P$ and $R^\nabla := \frac{1}{4} P^\top$. In particular, this choice ensures that $x_h \in (0,1)^{n_h}$.

As parameters, we found that $\kappa = 0.49$ and $\varepsilon = 10^{-3}$ performed well in our experiments. We applied multilevel gradient descent as described above. The images on the finest level had the resolution 1023×1023 and each level halfs the dimension until we reach 63×63. While we set $\lambda = 10^{-3}$, each λ_h was multiplied by two w.r.t. the next finer level. For instance, as 255×255 is the second coarsest level, we set $\lambda_h = \lambda \cdot 2^2$.

Using the setting defined above, we reconstructed the three images displayed in Figure 6.7. Explicit matrix vector multiplications to simulate tomographic projections are too costly on the fine grid, since we have to evaluate the objective function (6.60) and its gradient many times during the line search. Thus, we used the *astra-toolbox* [van16] with GPU acceleration [PBS11].

In our experiments, we observed that solving (6.66) until we reach convergence does not yield better coarse correction directions d^* than only perform two descent steps until $\psi_h(d^*) < 0$ holds. This choice is still sound to the theory developed in this chapter and reduces the computational effort to obtain a coarse correction.

Figure 6.7 illustrates that performing coarse corrections steps in between gradient descent steps accelerates the overall convergence.

6.5 Discussion

The chapter's basic idea was to use different discretization levels of an optimization problem to accelerate the convergence of gradient-based descent methods. To this end, we used the preliminary work of [Nas00] to develop a correction term, which computes a descent direction by utilizing a coarse discretization. We also studied the

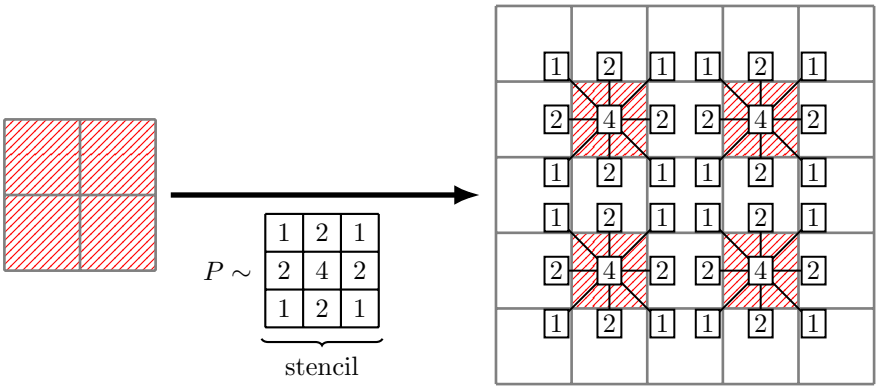


Figure 6.6: Using the *full weighting* prolongation map, we transfer each coarse grid point (left 2×2 grid) to a finer grid and distribute the values according to the stencil.

case of constrained optimization problems and found that we were able to use the same correction term and only had to treat the constraints with separate restriction maps.

For optimization on an open convex set in $\mathbb{R}^n$, we introduced concepts from differential geometry to employ gradient-based descent methods on manifolds. We defined a Riemannian structure on the open box $(l, u) \subseteq \mathbb{R}^n$ and proposed a method to transfer a retraction between manifolds.

Combined, we developed an algorithm that accelerates a gradient-based descent method using coarse corrections on multiple levels. Our experiments underpinned our theoretical findings.

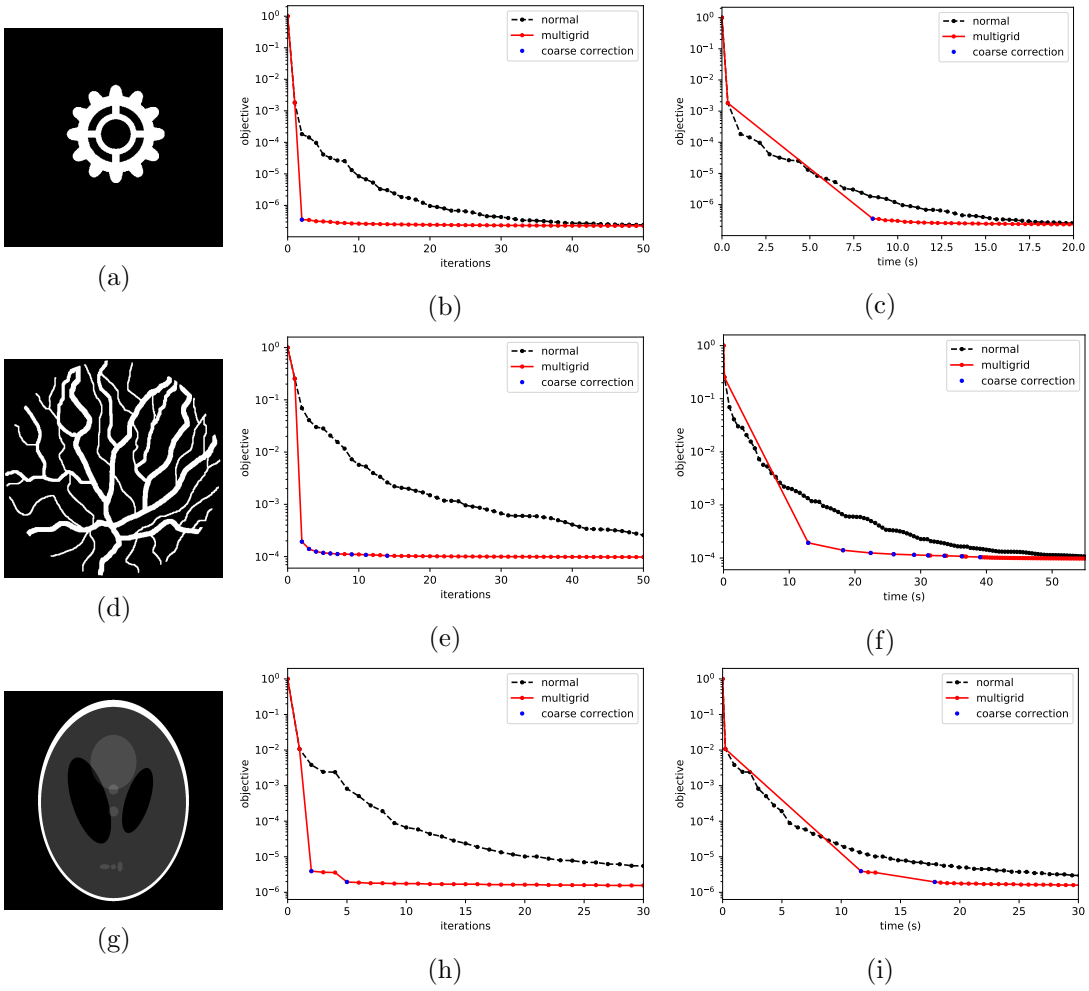


Figure 6.7: Each row represents a reconstruction of the image in the first column. In the second column we plotted the objective values during the iterations. We see that only few coarse correction steps (marked by blue dots) were needed to almost reach the optimal objective value. Comparing the time in the third column, the overhead due to the use of multilevel optimization is neglectable.

CHAPTER 7

Conclusion

In this thesis, we investigated the problem of reconstructing images from highly under-determined linear systems motivated by the field of *Discrete Tomography* (DT). As a solution of an under-determined linear system is never be unique, we considered constrained optimization to restrict the solution space according to prior knowledge. Throughout this thesis, we discussed two crucial features: the sparsity of the image gradient and the restriction of the domain of each pixel. The first property (sparse gradient) relates to the field of *Compressed Sensing* (CS) dealing with the unique reconstruction of sparse signals from incomplete linear measurements. However, results from CS mainly rely on choosing the linear system at random, making the direct application to DT impossible. Existing results for a deterministic choice lead to overly pessimistic recovery guarantees for DT.

The two main topics of this thesis were

(i) to guarantee unique recovery from insufficient projections, and

(ii) to construct optimization models that enforce sparsity of the image gradient or restrictions of the image domain.

Considering (i), in Chapter 3, we showed a performance bound to guarantee unique reconstruction for gradient sparse recovery via convex optimization. Hence, we contributed to CS by extending existing results for unbounded domains to bounded domains. Furthermore, we empirically validated that our theoretical result extend to tomographic projections. As a consequence, we exactly recover an image in DT only from few projections when knowing it has sparse gradient or lies in a bounded domain.

Additionally, in Chapter 5, we provided a lower bound for unique reconstruction when using non-convex optimization. Here, we showed that the number of connected components in an image is a crucial parameter in predicting the needed number of linear measurements, which was not considered before in the literature.

In view of (ii), in Chapter 4, we have proposed a novel convex relaxation and an accompanying algorithm for the non-binary discrete tomography problem. We theoretically and empirically showed that our novel relaxation is tighter than the relaxation commonly used.

As the number of connected components are crucial for unique reconstruction, we used in Chapter 5 integer programming to restrict the number of connected components in the resulting image. Moreover, we developed constraints to speed-up the convergence of branch and bound when solving the aforementioned integer programs.

Since the problem of Discrete Tomography origins from a variational problem, it naturally exists on different discretization levels. Therefore, in Chapter 6, we reviewed the field of *multilevel optimization* for constrained and unconstrained optimization. Viewing the image domain as a manifold and using tools from differential geometry and optimization on manifolds, we developed a first-order multilevel optimization algorithm. The developed multilevel algorithm did speed-up the convergence and allowed us to consider images of higher resolution.

Outlook On the theoretical side, we proved a feasible approximation of the lower bound to the number of needed measurements for unique reconstruction. While our empirical results showed that the approximation seems to be tight, theoretical evidence is still missing. In case of the one-dimensional finite difference operator this was done in [FM14; Zha16], however, it is not straightforward to extend it to the two-dimensional case.

We validated our numerical approaches on each of the proposed optimization models on carefully designed test instances and showed that they are tractable. Yet, on the algorithmic side, there is still space for improvement.

The convex relaxation from Chapter 4 is provably tighter than standard approaches. However, we were restricted to binary tomographic matrices. A next step would be to solve the subproblems according to non-binary projections, which is not straightforward.

In Chapter 5, we showed that the number of connected components is a good prior for exact reconstruction. However, solving the integer program for higher resolutions is time-consuming. Therefore, one needs to find efficiently solvable relaxations.

Chapter 6 showed that multilevel optimization applies to the reconstruction in discrete tomography. Although we proved that every coarse correction direction is a descent direction, it is still missing to predict or guarantee a minimum of descent in the objective function at the fine level. Such a result may enable us to prove convergence rates. Then, it would be possible to theoretically compare multilevel optimization against classical gradient descent.

Literature

[AB16] S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. 4th printing 2016. New York: Cambridge University Press, 2016 (page 53).

[ABB04] F. Alvarez, J. Bolte, and O. Brahic. "Hessian Riemannian Gradient Flows in Convex Programming". In: *SIAM Journal on Control and Optimization* 43.2 (2004), pp. 477–501 (pages 97, 102).

[Ame14] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. "Living on the Edge: Phase Transitions in Convex Programs with Random Data". In: *Information and Inference* 3.3 (2014), pp. 224–294 (pages 3, 29–32, 34, 38).

[AMS08] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton, N.J and Woodstock: Princeton University Press, 2008 (pages 4, 7–10, 88, 97–99).

[Ay17] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer. *Information Geometry*. Vol. 64. Cham: Springer International Publishing, 2017 (page 99).

[Bal09] E. Balas. "Integer Programming". In: *Encyclopedia of Optimization*. Ed. by C. A. Floudas and P. M. Pardalos. Boston, MA: Springer-Verlag, 2009, pp. 1617–1624 (pages 13, 64, 76).

[Bal18] E. Balas. *Disjunctive Programming*. Cham: Springer International Publishing, 2018 (pages 77, 78).

[Bar08] M. Bartholomew-Biggs. *Nonlinear Optimization with Engineering Applications*. Vol. 19. Boston, MA: Springer US, 2008 (page 44).

[Bat08] K. J. Batenburg. "A Network Flow Algorithm for Reconstructing Binary Images from Continuous X-rays". In: *Journal of Mathematical Imaging and Vision* 30.3 (2008), pp. 231–248 (page 59).

[BD09] T. Blumensath and M. E. Davies. "Sampling Theorems for Signals from the Union of Finite-dimensional Linear Subspaces". In: *IEEE Transactions on Information Theory* 55.4 (2009), pp. 1872–1882 (pages 3, 4, 14, 63, 65, 66).

[BDP17] S. Brocchi, P. Dulio, and S. M. C. Pagani. *Binary Tomography Reconstructions with Few Projections*. 2017 (page 1).

[Bel16] P. Belotti, P. Bonami, M. Fischetti, A. Lodi, M. Monaci, A. Nogales-Gómez, and D. Salvagnin. "On Handling Indicator Constraints in Mixed Integer Programming". In: *Computational Optimization and Applications* 65.3 (2016), pp. 545–566 (page 77).

[Bel54] R. Bellman. "The Theory of Dynamic Programming". In: *Bulletin of the American Mathematical Society* 60.6 (1954), pp. 503–516 (page 53).

[Ber15] D. P. Bertsekas. *Convex Optimization Algorithms*. Vol. 4. Optimization and computation series. Belmont, Massachusetts: Athena Scientific, 2015 (page 5).

[BHM00] W. L. Briggs, V. E. Henson, and S. F. McCormick. *A Multigrid Tutorial, Second Edition*. Society for Industrial and Applied Mathematics, 2000 (pages 87, 89).

[Bie96] D. Bienstock. "Computational Study of a Family of Mixed-integer Quadratic Programming Problems". In: *Mathematical Programming* 74.2 (1996), pp. 121–140 (pages 64, 69, 77).

[Bon08] P. Bonami, L. T. Biegler, A. R. Conn, G. Cornuéjols, I. E. Grossmann, C. D. Laird, J. Lee, A. Lodi, F. Margot, N. Sawaya, and A. Wächter. "An Algorithmic Framework for Convex Mixed Integer Nonlinear Programs". In: *Discrete Optimization* 5.2 (2008), pp. 186–204 (page 13).

[Bon15] P. Bonami, A. Lodi, A. Tramontani, and S. Wiese. "On Mathematical Programming with Indicator Constraints". In: *Mathematical Programming* 151.1 (2015), pp. 191–223 (pages 77, 79).

[Bou16] S. Bourguignon, J. Ninin, H. Carfantan, and M. Mongeau. "Exact Sparse Approximation Problems via Mixed-integer Programming: Formulations and Computational Performance". In: *IEEE Transactions on Signal Processing* 64.6 (2016), pp. 1405–1419 (pages 77, 79).

[Bra18] M. Branda, M. Bucher, M. Červinka, and A. Schwartz. "Convergence of a Scholtes-type Regularization Method for Cardinality-constrained Optimization Problems with an Application in Sparse Robust Portfolio Optimization". In: *Computational Optimization and Applications* 70.2 (2018), pp. 503–530 (pages 70, 75, 76).

[BS11] K. J. Batenburg and J. Sijbers. "DART: A Practical Reconstruction Algorithm for Discrete Tomography". In: *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 20.9 (2011), pp. 2542–2553 (page 1).

[Bus94] M. Bussieck, H. Hassler, G. J. Woeginger, and U. T. Zimmermann. "Fast Algorithms for the Maximum Convolution Problem". In: *Operations Research Letters* 15.3 (1994), pp. 133–141 (page 57).

[Buz08] T. M. Buzug. *Computed Tomography: From Photon Statistics to Modern Cone-beam Ct.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2008 (page 15).

[CDF19] M. Conforti, M. Di Summa, and Y. Faenza. "Balas Formulation for the Union of Polytopes is Optimal". In: *Mathematical Programming* 63.1 (2019), pp. 1–16 (page 78).

[Cha12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. "The Convex Geometry of Linear Inverse Problems". In: *Foundations of Computational Mathematics* 12.6 (2012), pp. 805–849 (page 30).

[Dak65] R. J. Dakin. "A Tree-search Algorithm for Mixed Integer Programming Problems". In: *The Computer Journal* 8.3 (1965), pp. 250–255 (pages 11, 64).

[Dem15] S. Dempe. *Bilevel Programming Problems: Theory, Algorithms and Applications to Energy Networks.* Energy systems. Heidelberg and New York: Springer, 2015 (pages 70, 73).

[Den14a] A. Denitiu, S. Petra, C. Schnörr, and C. Schnörr. "Phase Transitions and Cosparse Tomographic Recovery of Compound Solid Bodies from Few Projections". In: *Fundamenta Informaticae - Volume 135, issue 1-2 - Journals.* IOS Press, 2014, pp. 73–102 (pages 3, 15).

[Den14b] A. Deniţiu, S. Petra, C. Schnörr, and C. Schnörr. "An Entropic Perturbation Approach to TV-minimization for Limited-data Tomography". In: *DGCI 2014.* Ed. by E. Barcucci, A. Frosini, and S. Rinaldi. Lecture Notes in Computer Science. Cham: Springer, 2014, pp. 262–274 (page 1).

[Els18] J. Elstrodt. *Maß- und Integrationstheorie.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2018 (page 36).

[FHS16] R. Fukasawa, Q. He, and Y. Song. "A Disjunctive Convex Programming Approach to the Pollution-routing Problem". In: *Transportation Research Part B: Methodological* 94 (2016), pp. 61–79 (page 77).

[FM14] R. Foygel and L. Mackey. "Corrupted Sensing: Novel Guarantees for Separating Structured Signals". In: *IEEE Transactions on Information Theory* 60.2 (2014), pp. 1223–1247 (pages 38, 108).

[FQ13] J. Frikel and E. T. Quinto. "Characterization and Reduction of Artifacts in Limited Angle Tomography". In: *Inverse Problems* 29.12 (2013), p. 125007 (page 1).

[FR13] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing.* Applied and Numerical Harmonic Analysis. New York: Birkhäuser, 2013 (pages 2, 13, 19, 30).

[GGP99] R. J. Gardner, P. Gritzmann, and D. Prangenberg. "On the Computational Complexity of Reconstructing Lattice Sets from Their X-rays". In: *Discrete Mathematics* 202.1-3 (1999), pp. 45–71 (page 51).

[GHW20] P. Grohs, M. Holler, and A. Weinmann. *Handbook of Variational Methods for Nonlinear Geometric Data*. Cham: Springer International Publishing, 2020 (page 98).

[Gil16] J. C. Gilbert. "On the Solution Uniqueness Characterization in the L1 Norm and Polyhedral Gauge Recovery". In: *Journal of Optimization Theory and Applications* (2016), p. 31 (pages 20, 21).

[GMS05] P. E. Gill, W. Murray, and M. A. Saunders. "SNOPT: An SQP Algorithm for Large-scale Constrained Optimization". In: *SIAM Review* 47.1 (2005), pp. 99–131 (page 76).

[Gou13] E. Gouillart, F. Krzakala, M. Mézard, and L. Zdeborová. "Belief-propagation Reconstruction for Discrete Tomography". In: *Inverse Problems* 29.3 (2013), p. 035003 (pages 1, 51).

[Gra08] S. Gratton, M. Mouffe, P. L. Toint, and M. Weber-Mendonca. "A Recursive Formula-trust-region Method for Bound-constrained Nonlinear Optimization". In: *IMA Journal of Numerical Analysis* 28.4 (2008), pp. 827–861 (pages 87, 91, 92, 95, 96).

[GST08] S. Gratton, A. Sartenaer, and P. L. Toint. "Recursive Trust-region Methods for Multiscale Nonlinear Optimization". In: *SIAM Journal on Optimization* 19.1 (2008), pp. 414–444 (pages 87, 91, 92).

[GT10] S. Gratton and P. L. Toint. "Approximate Invariant Subspaces and Quasi-newton Optimization Methods". In: *Optimization Methods and Software* 25.4 (2010), pp. 507–529 (pages 87, 91).

[Her09] G. T. Herman. *Fundamentals of Computerized Tomography: Image Reconstruction from Projections*. 2. ed. Advances in pattern recognition. London: Springer, 2009 (pages 1, 15).

[HJ18] P. C. Hansen and J. S. Jørgensen. "Air Tools II: Algebraic Iterative Reconstruction Methods, Improved Implementation". In: *Numerical Algorithms* 79.1 (2018), pp. 107–137 (page 46).

[HK07] G. T. Herman and A. Kuba. *Advances in Discrete Tomography and Its Applications*. Applied and Numerical Harmonic Analysis. Boston, MA: Birkhäuser Boston, 2007 (pages 1, 15).

[HK99] G. T. Herman and A. Kuba, eds. *Discrete Tomography: Foundations, Algorithms, and Applications*. Applied and Numerical Harmonic Analysis. Boston, Mass.: Birkhäuser, 1999 (pages 1, 19).

[HL93a] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Vol. 305. Berlin, Heidelberg: Springer Berlin Heidelberg, 1993 (page 33).

[HL93b] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms Ii: Advanced Theory and Bundle Methods*. Vol. 306. Grundlehren der mathematischen Wissenschaften, A Series of Comprehensive Studies in Mathematics. Berlin and Heidelberg: Springer, 1993 (page 58).

[Hoj18] C. Hojny, I. Joormann, H. Lüthen, and M. Schmidt. *Mixed-integer Programming Techniques for the Connected Max-k-cut Problem*. 2018. URL: http://www.optimization-online.org/DB_HTML/2018/07/6738.html (page 81).

[HPZ16] V. Hovhannisyan, P. Parpas, and S. Zafeiriou. "Magma: Multilevel Accelerated Gradient Mirror Descent Algorithm for Large-scale Convex Composite Minimization". In: *SIAM Journal on Imaging Sciences* 9.4 (2016), pp. 1829–1857 (pages 87, 91).

[Hub64] P. J. Huber. "Robust Estimation of a Location Parameter". In: *The Annals of Mathematical Statistics* 35.1 (1964), pp. 73–101 (page 102).

[JH17] A. Javaherian and S. Holman. "A Multi-grid Iterative Method for Photoacoustic Tomography". In: *IEEE Transactions on Medical Imaging* 36.3 (2017), pp. 696–706. ISSN: 0278-0062 (pages 87, 91).

[JP07] S. Jokar and M. E. Pfetsch. *Exact and Approximate Sparse Solutions of Underdetermined Linear Equations*. 2007 (pages 77, 79).

[Kap15] J. H. Kappes, S. Petra, C. Schnörr, and M. Zisler. "Tomogc: Binary Tomography by Constrained Graphcuts". In: *Pattern recognition*. Ed. by J. Gall, P. Gehler, and B. Leibe. Vol. 9358. Lecture Notes in Computer Science. Springer, 2015, pp. 262–273 (pages 1, 51, 59).

[Kei17] S. Keiper, G. Kutyniok, D. G. Lee, and G. E. Pfander. "Compressed Sensing for Finite-valued Signals". In: *Linear Algebra and its Applications* 532 (2017), pp. 570–613 (pages 3, 51).

[KM16] M. Kočvara and S. Mohammed. "A First-order Multigrid Method for Bound-constrained Convex Optimization". In: *Optimization Methods and Software* 31.3 (2016), pp. 622–644 (pages 87, 91, 96).

[Kol15] V. Kolmogorov. "A New Look at Reweighted Message Passing". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.5 (2015), pp. 919–930 (page 52).

[KP19] J. Kuske and S. Petra. "Performance Bounds for Co-/sparse Box Constrained Signal Recovery". In: *Analele Universitatii "Ovidius" Constanta - Seria Matematica* 27.1 (2019), pp. 79–106 (page 4).

[KSP17] J. Kuske, P. Swoboda, and S. Petra. "A Novel Convex Relaxation for Non-binary Discrete Tomography". In: *SSVM 2017*. Ed. by F. Lauze, Y. Dong, and A. B. Dahl. Cham: Springer International Publishing, 2017, pp. 235–246 (pages 1, 4, 51).

[KV18] B. Korte and J. Vygen. *Combinatorial Optimization*. Vol. 21. Berlin, Heidelberg: Springer Berlin Heidelberg, 2018 (pages 10, 13, 45, 53, 64, 76, 81).

[LB16] T. Lukić and P. Balázs. "Binary Tomography Reconstruction Based on Shape Orientation". In: *Pattern Recognition Letters* 79 (2016), pp. 18–24 (page 1).

[LD08] Y. M. Lu and M. N. Do. "A Theory for Sampling Signals from a Union of Subspaces". In: *IEEE Transactions on Signal Processing* 56.6 (2008), pp. 2334–2345 (pages 3, 4, 14, 63, 65, 66).

[LD60] A. H. Land and A. G. Doig. "An Automatic Method of Solving Discrete Programming Problems". In: *Econometrica* 28.3 (1960), p. 497 (pages 11, 64).

[Lee12] J. M. Lee. *Introduction to Smooth Manifolds*. Vol. 218. New York, NY: Springer New York, 2012 (pages 4, 7–9, 88).

[Ley01] S. Leyffer. "Integrating SQP and Branch-and-bound for Mixed Integer Nonlinear Programming". In: *Computational Optimization and Applications* 18.3 (2001), pp. 295–309 (page 13).

[Li09] S. Z. Li. *Markov Random Field Modeling in Image Analysis*. 3. Aufl. Advances in pattern recognition. s.l.: Springer Verlag London Limited, 2009 (page 51).

[LPR96] Z.-q. Luo, J.-s. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge: Cambridge Univ. Press, 1996 (page 75).

[Nam13] S. Nam, M. E. Davies, M. Elad, and R. Gribonval. "The Cosparse Analysis Model and Algorithms". In: *Applied and Computational Harmonic Analysis* 34.1 (2013), pp. 30–56 (pages 26, 42, 64, 66, 67, 69, 70).

[Nas00] S. G. Nash. "A Multigrid Approach to Discretized Optimization Problems". In: *Optimization Methods and Software* 14.1-2 (2000), pp. 99–116 (pages 4, 87, 89, 91, 104).

[Nas10] S. G. Nash. *Convergence and Descent Properties for a Class of Multilevel Optimization Algorithms*. 2010 (pages 92, 94).

[Nas14] S. G. Nash. "Properties of a Class of Multilevel Optimization Algorithms for Equality-constrained Problems". In: *Optimization Methods and Software* 29.1 (2014), pp. 137–159 (pages 87, 91, 92, 94).

[Nat01] F. Natterer. *The Mathematics of Computerized Tomography*. Society for Industrial and Applied Mathematics, 2001 (pages 1, 15).

[Nat95] B. K. Natarajan. "Sparse Approximate Solutions to Linear Systems". In: *SIAM Journal on Computing* 24.2 (1995), pp. 227–234 (pages 14, 64, 69).

[NW06] J. Nocedal and S. J. Wright. *Numerical Optimization*. Second Edition. Springer Series in Operations Research and Financial Engineering. New York, NY: Springer Science+Business Media LLC, 2006 (pages 9, 89).

[NW13a] D. Needell and R. Ward. “Near-optimal Compressed Sensing Guarantees for Total Variation Minimization”. In: *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 22.10 (2013), pp. 3941–3949 (page 3).

[NW13b] D. Needell and R. Ward. “Stable Image Reconstruction Using Total Variation Minimization”. In: *SIAM Journal on Imaging Sciences* 6.2 (2013), pp. 1035–1058 (pages 3, 15).

[PBS11] W. J. Palenstijn, K. J. Batenburg, and J. Sijbers. “Performance Improvements for Iterative Electron Tomography Reconstruction Using Graphics Processing Units (gpus)”. In: *Journal of structural biology* 176.2 (2011), pp. 250–253 (page 104).

[Pot52] R. B. Potts. “Some Generalized Order-disorder Transformations”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 48.1 (1952), pp. 106–109 (page 63).

[PS14] S. Petra and C. Schnörr. “Average Case Recovery Analysis of Tomographic Compressive Sensing”. In: *Linear Algebra and its Applications* 441 (2014), pp. 168–198 (page 2).

[PSS13] S. Petra, C. Schnörr, and A. Schröder. “Critical Parameter Values and Reconstruction Properties of Discrete Tomography: Application to Experimental Fluid Dynamics”. In: *Fundamenta Informaticae* 125.3-4 (2013), pp. 285–312 (page 2).

[RLH14] S. Roux, H. Leclerc, and F. Hild. “Efficient Binary Tomographic Reconstruction”. In: *Journal of Mathematical Imaging and Vision* 49.2 (2014), pp. 335–351 (pages 17, 45, 46).

[Roc97] R. T. Rockafellar. *Convex Analysis*. Princeton paperbacks. Princeton, N.J: Princeton University Press, 1997 (pages 5, 22, 33, 35, 97).

[RW10] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Corr. 3rd print. Vol. 317. Grundlehren der mathematischen Wissenschaften. Berlin: Springer, 2010 (pages 5, 6, 34).

[Sch01] S. Scholtes. “Convergence Properties of a Regularization Scheme for Mathematical Programs with Complementarity Constraints”. In: *SIAM Journal on Optimization* 11.4 (2001), pp. 918–936 (page 76).

[Sch99] A. Schrijver. *Theory of Linear and Integer Programming*. Reprinted. Wiley-Interscience series in discrete mathematics and optimization. Chichester: Wiley, 1999 (pages 13, 64, 76).

[SGJ11] D. Sontag, A. Globerson, and T. Jaakkola. “Introduction to Dual Decomposition for Inference”. In: *Optimization for Machine Learning*. MIT Press, 2011 (pages 58, 60).

[SJP11] E. Y. Sidky, J. H. Jørgensen, and X. Pan. *Convex Optimization Problem Prototyping for Image Reconstruction in Computed Tomography with the Chambolle-pock Algorithm*. 2011 (page 1).

[Sto15] M. Storath, A. Weinmann, J. Frikel, and M. Unser. "Joint Image Reconstruction and Segmentation Using the Potts Model". In: *Inverse Problems* 31.2 (2015), p. 025003 (pages 64, 70).

[Tar12] D. Tarlow, K. Swersky, R. S. Zemel, R. P. Adams, and B. J. Frey. "Fast Exact Inference for Recursive Cardinality Models". In: *UAI*. 2012, pp. 825–834 (pages 53, 58).

[TP14] A. M. Tillmann and M. E. Pfetsch. "The Computational Complexity of the Restricted Isometry Property, the Nullspace Property, and Related Concepts in Compressed Sensing". In: *IEEE Transactions on Information Theory* 60.2 (2014), pp. 1248–1259 (page 14).

[Tro15] J. A. Tropp. "Convex Recovery of a Structured Signal from Independent Random Linear Measurements". In: *Sampling Theory, a Renaissance*. Ed. by G. E. Pfander. Cham: Springer International Publishing, 2015 (page 31).

[TSO01] U. Trottenberg, A. Schüller, and C. W. Oosterlee. *Multigrid*. San Diego: Academic Press, 2001 (pages 89, 103).

[Van14] R. J. Vanderbei. *Linear Programming: Foundations and Extensions*. 4. ed. Vol. 196. International series in operations research & management science. New York, NY: Springer, 2014 (pages 27, 51, 71, 73).

[van16] W. van Aarle, W. J. Palenstijn, J. Cant, E. Janssens, F. Bleichrodt, A. Dabravolski, J. de Beenhouwer, K. Joost Batenburg, and J. Sijbers. "Fast and Flexible X-ray Tomography Using the Astra Toolbox". In: *Optics express* 24.22 (2016), pp. 25129–25147 (page 104).

[Wer07] T. Werner. "A Linear Programming Approach to Max-sum Problem: A Review". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.7 (2007), pp. 1165–1179. ISSN: 0162-8828 (pages 55, 60).

[WG10] Z. Wen and D. Goldfarb. "A Line Search Multigrid Method for Large-scale Nonlinear Optimization". In: *SIAM Journal on Optimization* 20.3 (2010), pp. 1478–1503 (pages 87, 91).

[WS15] A. Weinmann and M. Storath. "Iterative Potts and Blake-zisserman Minimization for the Recovery of Functions with Discontinuities from Indirect Measurements". In: *Proceedings. Mathematical, physical, and engineering sciences* 471.2176 (2015), p. 20140638 (pages 63, 64, 70).

[WSH03] S. Weber, C. Schnorr, and J. Hornegger. "A Linear Programming Relaxation for Binary Tomography with Smoothness Priors". In: *Electronic Notes in Discrete Mathematics* 12 (2003), pp. 243–254 (pages 1, 59).

[Zha16] B. Zhang, W. Xu, J.-F. Cai, and L. Lai. “Precise Phase Transition of Total Variation Minimization”. In: *2016 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Piscataway, NJ: IEEE, 2016, pp. 4518–4522 (pages 28, 38, 108).

[Zis16] M. Zisler, S. Petra, C. Schnörr, and C. Schnörr. “Discrete Tomography by Continuous Multilabeling Subject to Projection Constraints”. In: *Pattern recognition*. Ed. by B. Rosenhahn and B. Andres. Vol. 9796. Lecture Notes in Computer Science. Springer, 2016, pp. 261–272 (pages 1, 59).

[ZK15] Y.-B. Zhao and M. Kočvara. “A New Computational Method for the Sparsest Solutions to Systems of Linear Equations”. In: *SIAM Journal on Optimization* 25.2 (2015), pp. 1110–1134 (pages 70, 73–75).

Index