hase retrieval with n*o*n-Euclidean Bregman based ge*o*metry

Jean-Jacques Godeme

Normandie Université

UNIVERSITÉ CAEN NORMANDIE

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité **MATHEMATIQUES**

Préparée au sein de l'**Université de Caen Normandie**

Phase retrieval with non-Euclidean Bregman based geometry

Présentée et soutenue par

JEAN-JACQUES GODEME

Thèse soutenue le 21/06/2024

devant le jury composé de :

M. MOHAMED JALAL FADILI	Professeur des universités - ENSICAEN	Directeur de thèse
M. PHILIPPE JAMING	Professeur des universités - UNIVERSITE BORDEAUX 1 SCIENCES ET TECHNOLOGIE	Président du jury
M. CLAUDE AMRA	Directeur de recherche au CNRS - Institut Fresnel (UMR 7249)	Membre du jury
MME CLARICE POON	assistant professeur - WARWICK - UNIVERITY OF WARWICK	Membre du jury
MME IRÈNE WALDSPURGER	Chargée de recherche - UNIVERSITE PARIS 9	Membre du jury
MME LAURE BLANC-FERAUD	Directeur de recherche au CNRS - UNIVERSITE NICE SOPHIA ANTIPOLIS	Rapporteur du jury
M. RUSSELL LUKE	Professeur - GOTTINGEN - GEORGE AUGUST UNIVERSITÄT	Rapporteur du jury

Thèse dirigée par **MOHAMED JALAL FADILI** (Groupe de recherche en informatique, image et instrumentation de Caen)

MIIΣ

GREYC
Laboratoire de recherche en sciences du numérique

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité MATHÉMATIQUES

Préparée au sein de l'Université de Caen Normandie

Phase Retrieval with Non-Euclidean Bregman based Geometry

Présentée et soutenue par
Jean-Jacques GODEME

Thèse soutenue publiquement le 21 06 2024 devant le jury composé de		
M. RUSSELL LUKE	Professeur, Universität Göttingen	Rapporteur du jury
M. LAURE BLANC-FÉRAUD	Directrice de recherche au CNRS, Université Côte d'Azur Nice-Sophia-Antipolis	Rapporteuse du jury
M. IRÈNE WALDSPURGER	Chargée de recherche au CNRS, Université Paris-Dauphine	Membre du jury
M. PHILIPPE JAMING	Professeur des universités, Université de Bordeaux	Membre du jury
M. CLARICE POON	Professeur, University of Warwick	Membre du jury
M. JALAL FADILI	Professeur des universités, ENSICAEN	Directeur de thèse
M. CLAUDE AMRA	Directeur de recherche au CNRS, Institut Fresnel	Membre du Jury

Thèse dirigée par JALAL FADILI, Groupe de recherche en informatique, image et instrumentation

UNIVERSITÉ CAEN NORMANDIE

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

MIIS

GREYC

To my grand-parents: Séverin (1918-2022), Eugénie (1945-2021) who passed away during my Ph.D.

To my parents and my close family for their unconditional love and support throughout. Akpe!

And to every math teacher and professor I had along the way.

"...il n'y a pas lieu de s'effrayer de l'évolution des mathématiques, et le refus de faire confiance à l'avenir ne cache le plus souvent que la paresse intellectuelle."

***J. Dieudonné**, Introduction fondements de l'analyse moderne.*

Acknowledgements

Le 18 juin 2020, à la fin du premier confinement, j'ai eu l'occasion de rencontrer Jalal pour la première fois lors d'un entretien en visioconférence. L'idée de faire une thèse après mon master 2 en optimisation à Paris-Saclay me paraissait encore improbable suite à tout ce qui s'est produit pendant cette année-là. Mais après cet entretien une lueur d'espoir est parue, ce qui me permet aujourd'hui d'écrire ces mots. En effet, j'ai compris que Jalal est une personne très enthousiaste, motivé, passionné et pédagogue. Je suis très respectueux envers lui tant pour ses qualités scientifiques qu'humaine. L'une de ses plus importantes qualités est sa patience envers moi malgré tous les aléas qui sont subvenu lors de ses quatre années qu'a duré ma thèse. J'ai eu la chance de rencontrer, non seulement un excellent directeur de thèse, mais également un mentor dans ma vie personnelle. Encore merci pour ses années et pour tout ce que tu as fait pour moi. Je ne l'oublierai jamais.

Je remercie Russell Luke et Laure Blanc-Féraud d'avoir accepté la tâche de rapporteurs pour la présente thèse. Merci également aux membres de ce jury, Irène Waldspurger, Philippe Jaming, Clarice Poon et Claude Amra. Avoir de si grands noms dans mon jury est un véritable honneur au vu du respect que chacun m'inspire.

Je dois tout spécialement remercier Claude Amra, Myriam Zerrad et Xavier Buet de l'Institut Fresnel de Marseille qui m'ont beaucoup aidé sur le côté expérimentale de cette thèse. Claude a pris sur lui à cause de mes lacunes en Optique ce qui m'a beaucoup appris sur l'intuition physique de nombreux concepts dit mathématiques, me permettant de construire un pont entre les deux disciplines là où se situe cette thèse. Tu es également une sorte de mentor pour moi en physique expérimentale.

Un certain nombre de chercheurs ont marqué ma courte expérience et m'ont permis de m'orienter vers la recherche.

À Paris-Saclay, j'aimerais particulièrement remercier Quentin Mérigot qui a été membre du comité de suivi de cette thèse et qui est le responsable de la formation master en Optimisation. Je te suis reconnaissant de m'avoir orienté vers Jalal lorsque j'étais un peu hésitant après mon M2. Je voudrais mentionner Léonard Monsaingeon et Thomas Gallouët pour mon stage de fin de master à l'INRIA sur le Transport Optimal.

Je pense à Guy Dègla de l'Institut de mathématiques et de Science Physique (IMSP) au Bénin qui m'a orienté vers l'optimisation au travers de son intérêt pour la discipline dès la fin de ma Licence. Après, je remercie Nadia Raïssi de l'Université Mohammed V de Rabat au Maroc qui m'a beaucoup appris durant mon stage. Les deux m'ont orienté et recommandé pour la bourse de Fondation Mathématiques Jacques Hadamard (FMJH) de Paris-Saclay.

Je voudrais aussi remercier les membres de l'équipe Image du GREYC qui m'ont accueilli pendant presque quatre années. Je voudrais remercier Arielle Perrette et Sophie Rastello, les administratifs, qui font tout leurs possibles pour alléger notre journée de chercheur en s'occupant de tâches non moins importantes.

Je voudrais mentionner mes "co-bureaux", tout du long, Antonio Silveti-Falls, Guillaume Jeanneret San Miguel et Rodrigo Maulen Soto qui ont du supporter tous mes humeurs, vous êtes de vrais collègues et plus.

Je souhaiterais mentionner les autres membres de l'équipe qui rendaient les pauses très intéressantes: Sébastien Fourey, Julien Rabin, Yukiko Kenmochi, Hamza Ennaji, Ryan Webster et tous les autres que je n'ai pu transcrire faute de place un grand merci. Un remerciement à Marjorie malgré nos dissensions, tu restes une personne intrigante. Je ne voudrais pas oublier la team Béhourd Matthieu et Raphaëlle , ce fut la découverte d'un sport intéressant.

Je voudrais remercier ceux qui me supportaient dans l'ombre. Mes amis de longues dates Joseph, Serge et notre nouveau membre Eloan. Vous avez toujours été là depuis que nous avons commencé les classes préparatoires. Vous m'avez prouvé à nombres reprises que dans les plus grandes difficultés, vous serez toujours disposés à me soutenir.

Je tiens à remercier les membres de l'association ASA que nous avons créés ensemble dans le but de faire de la vulgarisation sur notre continent : Branda, Clotilde, Cyprien, Danhane, Émile, Emmanuel, Florent, Ismaïl, Romziath. Grâce à nos séminaires du dimanche, j'explore d'autres champs des mathématiques.

Je voudrais aussi mentionner le séminaire du groupe de recherche MOP de Sarrebruck avec Peter Ochs qui m'ont tant appris.

Merci à mon père et ma mère qui ont tous donné pour m'enseigner que le travail est un trésor. Merci, à mon oncle pour support, lorsque j'étais souffrant et à tous ceux qui de près ont contribué à me maintenir en équilibre pour achever cette thèse. Merci à tous mes frères et ma sœur pour qui j'espère être un jour un modèle.

Jean-Jacques GODEME,
Caen, Avril 2024.

"Quelque chose de complexe n'est pas utile et tout ce qui est utile est simple ..."
M.K.

Abstract

In this work, we investigate the phase retrieval problem of real-valued signals in finite dimension, a challenge encountered across various scientific and engineering disciplines. It explores two complementary approaches: retrieval with and without regularization. In both settings, our work is focused on relaxing the Lipschitz-smoothness assumption generally required by first-order splitting algorithms, and which is not valid for phase retrieval cast as a minimization problem. The key idea here is to replace the Euclidean geometry by a non-Euclidean Bregman divergence associated to an appropriate kernel. We use a Bregman gradient/mirror descent algorithm with this divergence to solve the phase retrieval problem without regularization, and we show exact (up to a global sign) recovery both in a deterministic setting and with high probability for a sufficient number of random measurements (Gaussian and Coded Diffraction Patterns). Furthermore, we establish the robustness of this approach against small additive noise. Shifting to regularized phase retrieval, we first develop and analyze an Inertial Bregman Proximal Gradient algorithm for minimizing the sum of two functions in finite dimension, one of which is convex and possibly nonsmooth and the second is relatively smooth in the Bregman geometry. We provide both global and local convergence guarantees for this algorithm. Finally, we study noiseless and stable recovery of low complexity regularized phase retrieval. For this, we formulate the problem as the minimization of an objective functional involving a nonconvex smooth data fidelity term and a convex regularizer promoting solutions conforming to some notion of low complexity related to their nonsmoothness points. We establish conditions for exact and stable recovery and provide sample complexity bounds for random measurements to ensure that these conditions hold. These sample bounds depend on the low complexity of the signals to be recovered. Our new results allow to go far beyond the case of sparse phase retrieval.

Keywords: phase retrieval, inverse problems, stability to noise, inertial Bregman proximal gradient, partly smooth function, trap avoidance, variational regularization, sparsity, exact recovery, low complexity prior, robustness.

Résumé

Dans ce travail, nous nous intéressons au problème de reconstruction de phase de signaux à valeurs réelles en dimension finie, un défi rencontré dans de nombreuses disciplines scientifiques et d'inénierie. Nous explorons deux approches complémentaires : la reconstruction avec et sans régularisation. Dans les deux cas, notre travail se concentre sur la relaxation de l'hypothèse de Lipschitz-continuité généralement requise par les algorithmes de descente du premier ordre, et qui n'est pas valide pour la reconstruction de phase lorsqu'il formulée comme un problème de minimisation. L'idée clé ici est de remplacer la géométrie euclidienne par une divergence de Bregman non euclidienne associée à un noyau générateur approprié. Nous utilisons un algorithme de descente miroir ou de descente à la Bregman avec cette divergence pour résoudre le problème de reconstruction de phase sans régularisation. Nous démontrons des résultats de reconstruction exacte (à un signe global près) à la fois dans un cadre déterministe et avec une forte probabilité pour un nombre suffisant de mesures aléatoires (mesures Gaussiennes et pour des mesures structurées comme la diffraction codée). De plus, nous établissons la stabilité de cette approche vis-à-vis d'un bruit additif faible. En passant à la reconstruction de phase régularisée, nous développons et analysons d'abord un algorithme proximal inertiel à la Bregman pour minimiser la somme de deux fonctions, l'une étant convexe et potentiellement non lisse et la seconde étant relativement lisse dans la géométrie de Bregman. Nous fournissons des garanties de convergence à la fois globale et locale pour cet algorithme. Enfin, nous étudions la reconstruction sans bruit et

la stabilité du problème régularisé par un a priori de faible complexité. Pour celà, nous formulons le problème comme la minimisation d'une objective impliquant un terme d'attache aux données non convex et un terme de régularisation convexe favorisant les solutions conformes à une certaine notion de faible complexité. Nous établissons des conditions pour une reconstruction exacte et stable et fournissons des bornes sur le nombre de mesures aléatoires suffisants pour de garantir que ces conditions soient remplies. Ces bornes d'échantillonnage dépendent de la faible complexité des signaux à reconstruire. Ces résultats nouveaux permettent d'aller bien au-delà du cas de la reconstruction de phase parcimonieuse.

Mots-clés: reconstruction de phase, problème inverse, stabilité au bruit, algorithme du gradient proximal Bregman inertiel, fonction partiellement lisse, évitement de piège, parcimonie, a priori de faible complexité, robustesse.

Table of contents

Chapter 1

Introduction

Contents

1.1 Context and Motivations

This thesis studies a nonlinear, ill-posed inverse problem known as *phase retrieval*. It consists in recovering an arbitrary signal from the intensity of its linear measurements, *i.e.* from phaseless observations. In many applications, detectors or sensors have only access to the squared modulus of the Fresnel diffraction pattern of the radiation that is scattered from the object, leaving out desired structural information which comes from the phase or the sign of the signal. Historically, the first application of phase retrieval started with X-ray crystallography, and it now permeates many areas of imaging science with applications that include diffraction imaging, astronomical imaging, microscopy to name just a few; see the overviews [161, 97, 127] and references therein. One of the main applications motivating our work originates from optics precision. In this field, one is interested in characterizing the roughness of an optically polished surface. Often components (e.g., interference filters) exhibit optical losses of order 10^{-6} of the incident power. Super-polished surfaces are commonly used to circumvent this issue. Indeed, their roughness (responsible for losses by optical scattering) is very low compared to the illumination wavelength. Therefore, it is crucial to know how to characterize the roughness of polished surfaces. To do so, light scattering is ideal among the existing techniques because it is fast and noninvasive. The surface is illuminated with a laser source, and the diffusion is measured by moving a detector. Then the power spectral density of the surface topography can be directly measured thanks to the electromagnetic theory of light scattering; see [6, 7].

Our focus in this thesis will be on the case of real signals. Formally, suppose $\bar{x} \in \mathbb{R}^n$ is a signal

and that we are given information about the squared modulus of the inner product between $\bar{x}$ and m sensing/measurement vectors $(a_r)_{r\in[\![m]\!]}$. The noisy phase retrieval problem can be cast as:

$$\begin{cases} \text{Recover } \bar{x} \in \mathbb{R}^n \text{ from the measurements } y \in \mathbb{R}^m \\ y[r] = |a_r^* \bar{x}|^2 + \epsilon[r], \quad r \in [\![m]\!], \end{cases} \tag{GeneralPR}$$

where $[r]$ is the r-th entry of the corresponding vector, and , and $\epsilon \in \mathbb{R}^m$ models the noise during the acquisition process. Different types of noise can corrupt the measurements such as photon noise, thermal noise, Johnson noise. The measurement model (GeneralPR) is quite standard and is similar for instance to [55, 70, 63].

Since $\bar{x}$ is real-valued, the best one can hope for is to ensure that $\bar{x}$ is uniquely determined from its intensities up to a global sign. Phase retrieval is in fact an ill-posed inverse problem in general, and even for $\epsilon = 0$, checking whether a solution to (GeneralPR) exists or not is known to be NP complete [157]. The situation is even more challenging in presence of noise. Thus, one of the major difficulties is to design efficient recovery algorithms and find conditions on m, $(a_r)_{r\in[\![m]\!]}$ and ϵ which guarantee exact (up to a global sign change) and robust recovery.

1.2 Prior Work

Our review here on the phase retrieval problem and how to solve it, is by no means exhaustive and the interested reader should refer to the following references for comprehensive reviews [161, 97, 78, 175, 127].

1.2.1 Phase retrieval without regularization

Feasibility formulation of constrained phase retrieval In the one-dimensional, on the continuum setting with Fourier measurements, it was shown by [1, 2, 181] (see also [28, 32] in the discrete case) that the phase retrieval problem without any prior constraints lacks uniqueness (up to trivial ambiguities). This fundamental barrier does not apply in higher dimensions as pointed out in [190] and shown in [93] for band-limited 2D signals, and uniqueness was shown to hold "generically" in [19].

To circumvent this barrier, workarounds have been proposed that involve adding a constraint either implicitly or explicitly. Phase retrieval is then formulated as a feasibility problem, that is, as finding some point in the intersection of the set of points satisfying the constraints implied by the data measurements in (GeneralPR), and the set of points satisfying constraints expressing some prior knowledge on the object to recover, such as support, band-limitedness, non-negativity, sparsity, etc. The Gerchberg and Saxton algorithm [82], proposed in the early 70's in the optics literature, is an alternating projection algorithm to solve such a feasibility problem. Improved variants include Fienup's basic input-output and the hybrid input-output (HIO) [65, 79, 66]. For the case of a support constraint alone, it has been identified by [23] that HIO corresponds to the now well-known Douglas-Rachford algorithm. Other fixed-point iterations based on projections that apply to constrained phase retrieval have also been proposed, such as the HPR scheme [24], or RAAR [125] which is a relaxation of Douglas-Rachford. Thanks to a wealth of results in the variational analysis community, some convergence properties of these algorithms for the phase retrieval problem are now known. One has to distinguish between the two important cases for feasibility problems: consistent and inconsistent.

For consistent phase retrieval problems, it is known for instance that alternating projections is locally linearly convergent at points of intersection provided that the constraints do not intersect tangentially [73, 25, 142, 126, 112, 113]. Similar results are also known for Douglas-Rachford [94, 150]. Global convergence guarantees are however only conjectured, and translating the nontangential intersection into conditions on m and $(a_r)_{r\in[\![m]\!]}$ remains open.

For the inconsistent case, it was argued in [127] that almost any constraint, in particular compact support, will be inconsistent with the measurement process in optical phase retrieval problems. This means that the corresponding feasibility problems are inconsistent. In this even more challenging inconsistent phase retrieval setting, the only two works that we are aware of where local linear convergence of alternating projections and relaxed Douglas-Rachford to local best approximation points is established are [129, Theorem 3.2 and Example 3.6] and [128, Theorem 4.11 and Section 5].

Unconstrained phase retrieval In the unconstrained setting of (GeneralPR), the dominant approach in computational phase retrieval is to use oversampling, i.e. to take more measurements, in order to ensure well-posedness and improve the performance of phase retrieval algorithms. This idea of oversampling has been known for a while, and for instance in non-crystallographic modalities [134]. From a theoretical point of view, for the case where $(a_r)_{r\in[\![m]\!]}$ is a frame (redundant complete system), the authors in [16, 15] derived various necessary and sufficient conditions for the uniqueness of the solution, as well as algebraic polynomial-time numerical algorithms valid for very specific choices of $(a_r)_{r\in[\![m]\!]}$. This approach is however of theoretical interest only and has some drawbacks, for instance that it requires specific types of measurements that cannot be realized in most applications of interest.

A very different route, which is now an established approach in the applied mathematics literature to understand fundamental limits of phase retrieval and its stability, consists in assuming that $(a_r)_{r\in[\![m]\!]}$ are sampled from an appropriate distribution and using probabilistic arguments to get lower bounds on oversampling (i.e. m/n) and on ϵ to ensure exact (up to sign or phase change in the complex case) and stable recovery with high probability. This can be done either through convex relaxation or by directly attacking the nonconvex formulation of the phase retrieval problem;see [55, 70, 164, 63]. The recovery error and oversampling bounds derived in those papers can be compared to the fundamental lower bounds established in [17, 77, 63]. We would like to mention that we are aware that this setting might not always be realistic from an application perspective as it may sometimes involve changing the data measurements, imposed by the physical imaging system, to fit the theory. Nonetheless, this still makes sense in some applications such as the one motivating our work (precision in optics), where the CDP (Coded Diffraction Patterns) measurement model can be implemented.

Convex relaxation The key ingredient is to use a well-known trick turning a quadratic function on $\mathbb{R}^n$, such as in the data measurement mapping in (GeneralPR), into a linear function on the space of $n \times n$ matrices [31, 84]. Thus the recovery of a vector from quadratic measurements is lifted into that of recovering a rank-one Hermitian semidefinite positive (SDP) matrix from affine constraints, and the rank-one constraint is then relaxed into a convenient convex one. The two most popular methods in this line are PhaseLift [56, 55] and PhaseCut [180]. Both approaches are inspired by the matrix completion problem [54] and they differ in the way factorization takes place. Exact and robust recovery with random Gaussian or CDP (Coded Diffraction Patterns) measurements using PhaseLift was established in [55, 51, 52]. For Gaussian measurements, [55] showed that exact recovery by PhaseLift holds for a sampling complexity bound $m \gtrsim n\log(n)$. This has then been improved in [51] to a universal result with $m \gtrsim n$. Exact recovery by PhaseLift for CDP measurements was established in [52] for $m \gtrsim n\log^4(n)$, and has been improved to $m \gtrsim n\log^2(n)$ in [87]. While SDP based relaxations lead to solving tractable convex problems, the prospect of squaring the number of unknowns make them computationally prohibitive and impractical as n increases. Since then, there has been a resurgence in the proposal of direct nonconvex methods.

Nonconvex formulations The general strategy here is to use an initialization technique that lands one in a neighborhood of the optimal solution (up to global sign or phase change) where a usual iterative procedure from nonlinear programming with carefully chosen parameters can perform reliably.

In [53], the authors use a spectral initialization and propose a gradient-descent type algorithm (Wirtinger flow) for solving the general complex phase retrieval problem by casting it as

$$\min_{z\in\mathbb{C}^n} f(z) \stackrel{\text{def}}{=} \frac{1}{4m}\sum_{r=1}^{m}\left(y[r]-|a_r^* z|^2\right)^2. \tag{1.2.1}$$

For an appropriate (Wirtinger) gradient-descent step-size, they showed that with high probability, the scheme converges linearly to the true vector (up to a global phase change) for both Gaussian and CDP measurements provided that m is on the order of n up to polylogarithmic terms.

A truncated version of the Wirtinger flow was proposed in [63] which uses careful selection rules providing a tighter initial guess, better descent directions and step-sizes, and thus enhanced performance. For Gaussian measurements, truncated Wirtinger flow was also shown to converge linearly to the correct solution and is robust to noise provided that $m \gtrsim n$. Other variants of Wirtinger flow possibly and/or other initializations were proposed in [192] and [182], and were shown to enjoy similar guarantees in the noiseless case for Gaussian measurements. The Polyak subgradient method to minimize $\frac{1}{m}\sum_{r=1}^{m}|y[r]-|a_r{}^\top z|^2|$ on $\mathbb{R}^n$ was proposed and analyzed in [69] for noiseless real phase retrieval with real isotropic sub-Gaussian measurements. When properly initialized, its linear convergence was also shown for $m \gtrsim n$.

An alternating minimization strategy, alternating between phase update and vector update, with a resampling-based initialization has been proposed in [141] and was shown to enjoy noiseless exact recovery for $m \gtrsim n\log(n)^3$. A truncated version of the spectral initialization followed by alternating projection was also proposed in [179] with exact recovery guarantees for Gaussian measurements under the sample complexity bound $m \gtrsim n$.

The authors in [168] studied the landscape geometry of the nonconvex objective in (1.2.1) for Gaussian measurements. They showed that for large enough number of measurements, i.e., $m \gtrsim n\log(n)^3$, there are no spurious local minimizers, all global minimizers are equal to the correct signal $\bar{x}$, up to a global sign or phase, and the objective function has a negative directional curvature around each saddle point (that we coin strict saddles in this manuscript). This allowed them to describe and analyze a second-order trust-region algorithm to find a global minimizer without special initialization. The work of [64] provides an analysis of global convergence properties of gradient descent for Gaussian measurements and heavily relies on Gaussianity of the initialization. They required a sample complexity bound $m \gtrsim n\text{poly}\log(m)$ without making explicit the linear local convergence rate.

Stability to noise In phase retrieval, understanding the impact of noise is crucial because real-world measurements are invariably corrupted by it. Thus, establishing stability of phase retrieval to (small enough) noise is of paramount importance. For convex relaxations, Candès and Li showed in [51] that a noise-aware variant of PhaseLift is stable against additive noise with a reconstruction error bound $O\left(\|\bar{x}\|, \frac{\|\epsilon\|_1}{m\|\bar{x}\|}\right)$ as soon as $m \gtrsim n$ (complex) Gaussian measurements are taken (see also [55] where the sample complexity was $m \gtrsim n\log(n)$). This is of course only meaningful if the signal-to-noise ration is sufficiently high. This result has been extended to the case of sub-Gaussian measurements in [102]. For nonconvex formulations, Huang and Xu in [132] analyzed the performance of the Wirtinger flow and showed that any solution of this algorithm enjoys a reconstruction error upper bound $O\left(\min\left\{\frac{\sqrt{\|\epsilon\|}}{m^{1/4}}, \frac{\|\epsilon\|}{\|\bar{x}\|\sqrt{m}}\right\}\right)$ as soon as $m \gtrsim n$. The amplitude and the reshaped Wirtinger flow algorithms are stable against additive noise as shown respectively in [192],[182] and [80]. Indeed, these authors showed that the reconstruction error scales as $O\left(\frac{\|\epsilon\|}{\sqrt{m}}\right)$. The convergence result is obtained under the specific assumption that $\|\epsilon\|_\infty \lesssim \|\bar{x}\|$. In [188], the authors study the performance of the amplitude-based model and showed the solution satisfies the following reconstruction upper bound $O\left(\frac{\|\epsilon\|}{\sqrt{m}}\right)$ as soon as $m \gtrsim n$. The truncated Wirtinger flow [63], which can account even for Poisson noise, has been shown to be stable with a reconstruction error bound that scales as $O\left(\frac{\|\epsilon\|}{\sqrt{m}\|\bar{x}\|}\right)$

under the assumption $\|\epsilon\|_\infty \lesssim \|\bar{x}\|^2$ provided that $m \gtrsim n$. It was shown there that this is the best statistical guarantee any algorithm can achieve by deriving a fundamental lower bound on the minimax estimation error.

1.2.2 Phase retrieval with regularization

As argued above, (GeneralPR) is a severely ill-posed inverse problem in general unless either oversampling or some prior knowledge is available on the underlying signal $\bar{x}$. The use of a well-chosen prior allows to restrict the inversion process to an appropriate subset of $\mathbb{R}^n$ containing the plausible solutions including $\overline{\mathcal{X}}$. In turn, this allows to reach the land of well-posedness. A standard way to implement this idea consists in adopting a variational framework where the sought-after solutions are those where a prior penalty/regularization function R is the smallest. This approach is in line with variational regularization theory pioneered by Tikhonov [170]. Put formally, this amounts to solving the following optimization problem

$$\inf_{x\in\mathbb{R}^n} \left\{ F_{y,\lambda}(x) \overset{\text{def}}{=} F(x) + G(x) = \left\| y - |Ax|^2 \right\|^2 + \lambda R(x) \right\}, \tag{$\mathcal{P}_{y,\lambda}$}$$

where $A = [a_1, \dots, a_m]^\top$ and $R : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function which is intended to promote objects similar or close to $\bar{x}$. $\lambda > 0$ is the regularization parameter which balances the trade-off between fidelity and regularization. It is immediate that F is $C^2(\mathbb{R}^n)$ but is nonconvex due to the quadratic measurements. Besides, his gradient is not Lipschitz continuous. In this setting, we can associate to the objective the following function or kernel

$$\psi(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2 . \tag{1.2.2}$$

$\psi \in C^2(\mathbb{R}^n)$ has full domain and 1−strongly convex function with a gradient that is Lipschitz over bounded subsets of $\mathbb{R}^n$. It turns out that F is smooth relative to ψ (see Section 2.3 for the definition and discussion of the notion of relative smoothness). Therefore, the problem ($\mathcal{P}_{y,\lambda}$) is amenable to the efficient Bregman proximal gradient scheme; see Chapter 5 for a detailed description and discussion.

1.2.2.1 Algorithms to solve ($\mathcal{P}_{y,\lambda}$) and their guarantees

Inertial Bregman proximal gradient algorithms Inertial methods emerged from the quest of accelerating the convergence of first-order optimization methods such as gradient descent. This starts with the Heavy-ball with friction [151] method for gradient descent ($G \equiv 0$). This approach can be interpreted as a discretization of a nonlinear second-order dynamical system, specifically an oscillator with viscous damping. This idea permeates now all the optimization techniques and has been applied for instance to the proximal point method [3, 4] and to the inertial Forward-Backwards type methods [135, 10, 123]. In terms of acceleration for convex programming, the accelerated FISTA method [27, 140] achieves a convergence rate of $O(k^{-2})$ for the sequence of objective functions which has been improved to $o(k^{-2})$ in [9], with convergent iterates in [58].

For Bregman-based methods, first-order methods achieve the convergence $O(k^{-1})$ for the sequence of objective [34, 21, 40, 124, 169] as in the Euclidean case. A natural question is whether the Bregman proximal gradient algorithm can be accelerated in the relative smooth setting. This question has been raised in several works, including [21, 124], and the survey paper [169, Section 6]. Positive answers have already been provided under somewhat strict additional regularity assumptions. When the entropy ψ is a strongly convex Legendre kernel and the smooth part of the objective has a Lipschitz continuous gradient, the Improved Interior Gradient Algorithm in [13] admits an accelerated $O(1/k^2)$ convergence rate on the objective, by using the same inertial technique as Nesterov-type methods. For a subclass of relatively smooth functions, [89] shows that the convergence rate of the objective can be improved from

to $O(1/k^\kappa)$ where $\kappa \in [1, 2]$ is determined by some crucial triangle scaling property of the Bregman distance, whose genericity is unclear. The general case was still open until the work of [72] which showed that the $O(1/k)$ rate is optimal for first-order algorithms over the class of relatively smooth functions, and this cannot be improved in general. As far as the nonconvex case is concerned, inertial versions of the proximal gradient method were analyzed in [136, 185, 95, 187] when the entropy is strongly convex. All these works use either backtracking or line search on both the extrapolation (inertial) parameter and the descent step-size. Global convergence of the iterates under KL were proved in [136, 187] while [185] showed linear convergence under certain error bound condition.

Activity identification Finite activity identification of underlying manifolds is an important phenomenon that occurs when different types of algorithms are used to solve structured nonsmooth minimization problems. Such analysis can be traced back to the work of [48] and [76], where a the non-degeneracy condition is used to ensure that the optimal active constraints are identified after finitely many iterations. They showed how polyhedral faces of convex sets could be identified finitely. Later [186] extended these results by introducing the concept of smooth identifiable surface and providing an algorithm that identifies the active constraints in convex problems in a finite number of iterations. This work was then generalized in [111] to the notion of partial smoothness for general non necessarily convex problems.Among algorithms that identify active manifolds of partly smooth functions, one can cite the (sub)gradient projection method, Newton-like methods, and the proximal point algorithm as shown in [92, 91, 90] have shown for the (sub)gradient projection method, Newton-like methods, and the proximal point algorithm. A comprehensive study of finite activity identification as well as sharp local linear convergence has been established by Liang et al. in a series of papers for a variety of operator splitting algorithms: forward-backward-type algorithms including accelerated ones [116, 118, 117], for Douglas-Rashford/ADMM [120] and for the Primal-Dual splitting [119].

Strict saddle avoidance A driving theme in nonconvex optimization, supported by empirical evidence, is that simple algorithms often work well in highly nonconvex and even nonsmooth settings by avoiding "bad" critical points. A growing body of literature provides one compelling explanation for this escape property or trap avoidance phenomenon. Namely, typical smooth objective functions provably satisfy the strict saddle property, meaning each critical point is either a local minimizer or has a direction of strictly negative curvature. For such functions, either randomly initialized gradient-type methods [86, 146], or stochastically perturbed gradient methods [149, 44] provably escape all strict saddle points, generically on initialization or on the noise. At the heart of the analysis in all these work is the use of the Center Stable Manifold Theorem [162, Theorem III.7] which finds its roots in the work of Poincaré. In [86], it was proved that the heavy ball method with friction, applied to a C^2 Morse function, provably converges to a local minimizer generically on initialization. Morse functions are known to be generic in the Baire sense in the space of C^2 functions (see [11]). In [110], it has been shown that when the objective function is sufficiently smooth with a Lipschitz continuous gradient, a large class of first-order methods avoid strict saddle points. In the nonsmooth setting, Euclidean proximal methods as explored in [68], are effective in avoiding a nonsmooth version of strict saddle points: "active" strict saddle points, which are strict saddle points with respect to an underlying activity manifold. In parallel to [68], we would like to mention the recent work of [33] which shows that stochastic subgradient descent also escapes "active" strict saddle points. The last two papers rely on important geometric conditions that turn out to be generic in the space of tame weakly convex functions.

1.2.2.2 Recovery and stability guarantees for $(\mathcal{P}_{y,\lambda})$

Regularized phase retrieval is an active area of research. Our review of this problem is not exhaustive and readers interested in a comprehensive and extended overview should refer to the following references [161, 97, 175, 127, 78].

Sparse phase retrieval When the signal of interest is $s-$sparse w.r.t some basis and and the goal is to recover the signal from a few measurements $m \ll n$, this problem is referred as "compressive or sparse phase retrieval". From a theoretical perspective, generic sensing vectors $(a_r)_{r\in[\![m]\!]}$ are injective (up to a global sign change) in the class of real $s-$sparse signals as soon as the number of measurements satisfies $m \geq 2s-1$ [184]. We recall that the natural information theoretical lower-bound is $m \gtrsim s\log(n)$ for solving the problem using any approach. Whereas for Gaussian sensing vectors, [143] show that $m \gtrsim s\log(en/s)$ separate signals well. In [178], the authors introduced a notion of strong Restricted Injectivity Property (s-RIP) which holds for the class of Gaussian sensing vectors and they showed that solving $(\mathcal{P}_{\bar{y},0})$ when R is the ℓ_0-norm is equivalent to solving the same problem replacing ℓ_0 with the ℓ_1 norm for sensing vectors satisfying the s-RIP. For Gaussian sensing vectors, the latter holds for $m \gtrsim s\log(en/s)$. Stable sparse phase retrieval under the s-RIP was studied in [80]. Other works in the same vein include [131, 189, 80, 184, 159, 98, 99, 100, 18].

We can categorize the methods to solve the sparse phase retrieval problem into three groups. The first considers convex relaxation, the second tackles directly the nonconvex problem and the third manually designs the measurement vectors.

Again, lifting methods such as the PhaseLift or PhaseCut[1] can be used to convexify the constraint in $(\mathcal{P}_{y,\lambda})$ while sparsity on the lifted rand-one matrix is now to be promoted entry-wise or on rows. This regularization entails that the rank-one matrix to be recovered is s^2 sparse and thus, as expected, the sample complexity for exact recovery from Gaussian measurements is $m \gtrsim s^2\log(n)$ [114]. However, this problem becomes less tractable and it is not possible to achieve the natural theoretical lower-bound using this approach [114, 144]. Another approach in this setting is PhaseMax [88] which consists in relaxing the nonconvex constraint set in $(\mathcal{P}_{\bar{y},0})$ from equality to inequality (*i.e.* from the sphere constraint to the ball one), and then to solve the resulting linear program. This method achieves the optimal sample complexity $m \gtrsim s\log(n/s)$ for Gaussian sensing vectors. However, it requires an anchor or initialization that is sufficiently correlated with the true signal which requires $m \gtrsim s^2\log(n)$ to be successful. In [133], the authors use a convex relaxation and propose an atomic norm that favors low-rank and sparse matrices. They achieve nearly optimal sample complexity *i.e.* bound $m \gtrsim s\log(en/s)$. Regarding the stability of the reconstruction against additive noise, the same authors showed that the sparse with low-rank atomic norm regularization achieves a reconstruction error bound of $O(\sigma\sqrt{\frac{s}{m}\log(en/s)})$ where σ is the standard deviation.

Concerning methods that study directly the sparse phase retrieval problem, it has been shown that $m \gtrsim s^2\log(n)$ are sufficient to provably recover the original vector (up to global sign/phase change) [50, 141, 183, 191, 101]. The authors in [50] proposed a method to find a good initialization of the problem which requires that $m \gtrsim s^2\log(n)$. The authors in [141] proposed an alternating minimization strategy to reconstruct the signal. Sparta [183] uses an amplitude-based instead of an intensity-based measurement which is clearly nonsmooth and [191] proposed a sparse version of the classical Wirtinger flow [53]. The authors in [101] proposed the Copram which combines Alternating minimization and the Cosamp [137], and they showed that reconstruction is possible with $O(s^2\log(n))$ measurements. In the general case of block sparsity or group Lasso, they showed that exact recovery (up to a global sign change) is possible with $m \gtrsim \frac{s^2}{B}\log(n)$ where B is the size number of the blocks slightly improving the bound on the number of measurements. As far as robust recovery is concerned, it was proved

[1] Even though, we are not aware of any work adapting the PhaseCut to sparse phase retrieval.

in [50, 189] that solving ($\mathcal{P}_{y,\lambda}$) achieves a reconstruction error bound of $O(\sigma\sqrt{(s/m)\log(n)})$ which is very close to the optimal rate for the classic compress sensing.

The third class of methods that design the sensing vectors usually achieves near-optimal sample complexity bounds, we refer to [14, 148] but they are of limited interest for our work.

General regularized phase retrieval As reviewed above, most existing work focuses on the recovery of sparse signals from phaseless measurements. On the other hand, real signals and images involve much richer structure and complexity such as being piecewise smooth. In this case, a wise choice of the regularizer would the is the popular Total Variation (TV) seminorm, or sparsity in some frame. This scope is quite recent for the phase retrieval. For the TV phase retrieval, we refer to [29, 30]. In [29], the authors combined the standard Fienup's Hybrid input-output [79] method that is well-known to be the Douglas-Rachford [23] with TV regularization based on a primal-dual method. This was applied to optical diffraction tomography and the sensing vectors are the Non-Uniform Fourier Transform. In [30], they extend the scope to moving objects. See also [147] for an algorithmic framework based on Fienup methods with general semialgebraic regularizers.

In the general setting, we have to cite the work in [107], where the authors consider the reconstruction of a real vector living in a general constraint subset of $\mathbb{R}^n$ from sub-Gaussian measurements. They showed that Empirical Risk Minimization to solve the noisy phase retrieval produces a signal close enough to the true signal up to sign change and this error depends on the Gaussian width of the subset and the signal-to-noise ratio of the problem. Phase retrieval with general regularization is studied in [165], where the authors showed that the main problem for achieving the optimal sample complexity is the initialization step. However, it is still an open question to find a good strategy to find an anchor or initialization that is close enough or sufficiently correlated with the true vector beyond the sparse case, and with a reasonable bound on the number of measurements, i.e. that does not scale as the square of the intrinsic dimension of the vector to recover.

1.3 Contributions

We recall that throughout this manuscript, we consider that the signal is real valued [2]. Our work makes contributions in two distinct areas: first, by exploring the problem of phase retrieval without any prior knowledge on the signal that we aim to recover, and second by delving into the case of phase retrieval with regularization.

1.3.1 Phase retrieval without regularization

In this part of the thesis, we formulate (GeneralPR) as the minimization problem (1.2.1). Inspired by [40], we propose a mirror descent (or Bregman gradient descent) algorithm with backtracking associated to a wisely chosen Bregman divergence, hence removing the classical global Lipschitz continuity requirement on the gradient of the nonconvex objective (see (1.2.1)).

In the deterministic case without any specific structural assumption on the sensing vectors, we show that for almost all initializers, mirror descent converges to a critical point near the true vectors (up to sign ambiguity) where the objective has no direction of negative curvature, *i.e.*, a critical point which is not a strict saddle point. We show that mirror descent to solve (1.2.1) is stable against additive noise and this in turn provides recovery error bounds of the noisy phase retrieval problem (GeneralPR). Theses results are summarized in the following theorem. Let us denote $\overline{\mathcal{X}} = \{\pm\bar{x}\}$.

Theorem 1.3.1 (Exact and stable recovery guarantees for deterministic measurements).

[2] This is motivated by main application in light scattering where the roughness of a surface to be recovered is real.

Consider the phase retrieval problem cast as (1.2.1). Let $(x_k)_{k\in\mathbb{N}}$ be a bounded sequence generated by Algorithm 3. Then,

(i) the sequence $(x_k)_{k\in\mathbb{N}}$ has a finite length, converges to a critical point and the values $(f(x_k))_{k\in\mathbb{N}}$ are nonincreasing.
For constant descent step-sizes and without backtracking then,

(ii) for Lebesgue almost all initializers x_0, the sequence $(x_k)_{k\in\mathbb{N}}$ converges to a critical point which cannot be a strict saddle.

(iii) Assume that $\operatorname{Argmin}(f) \neq \emptyset$. Let $\rho, \sigma > 0$ such that $\rho > \frac{\sqrt{2}\|\epsilon\|}{\sqrt{m\sigma}}$ and define the radius $r \leq \sqrt{\frac{\rho^2 - \frac{2\|\epsilon\|^2}{m\sigma}}{\max(\Theta(\rho),1)}}$ where $\Theta(\cdot)$ is a function to be specified later. If the initial point $x_0 \in B(\overline{\mathcal{X}}, r)$ and f is σ-strongly convex relative to ψ on $B(\overline{\mathcal{X}}, \rho)$ then we have

$$\forall k \in \mathbb{N}, \quad x_k \in B(\overline{\mathcal{X}}, \rho) \text{ and } \operatorname{dist}^2(x_k, \overline{\mathcal{X}}) \leq (1-\gamma\sigma)^{k-1}\rho^2 + 2\frac{\|\epsilon\|^2}{m\sigma}. \tag{1.3.1}$$

The deterministic stable recovery results of Theorem 1.3.1 require for instance a local relative strong convexity condition as well as a good enough initial guess. A natural question to ask is when these conditions are verified. In turns out that this is indeed the case in the oversampling regime with random measurements drawn from appropriate random ensembles. We consider two random measurements models: i.i.d Gaussian measurements and the CDP model.

For Gaussian measurements, and in the regime where the signal-to-noise ratio is large enough, we provide a complete geometric characterization of the landscape of the nonconvex objective provided that $m \gtrsim n\log^3(n)$. In turn, this allows us to describe the set of the critical points of f as the union of the strict saddle points and global minimizers of f. From this, we provide a generic convergence result of our algorithm to a point in $\operatorname{Argmin}(f)$, which is near the true vector (up to sign ambiguity), as soon as the number of samples is large enough. If $m \gtrsim n\log(n)$, using a spectral initialization method, we provide a local convergence to a vector in the neighborhood of the target vector (up to sign ambiguity). Our main result for Gaussian measurements is the following.

Theorem 1.3.2 (Exact and stable recovery guarantees for Gaussian measurements). *Fix $\lambda \in \left]\frac{1}{9\sqrt{2}}, 1\right[$, ϱ, ς and $\widetilde{\epsilon}$ as in Theorem 4.3.2. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 3.*

(i) If the number of measurements m is large enough, i.e., $m \geq C(\varrho)n\log^3(n)$, then for almost all initializers x_0 of Algorithm 3 with the step-size $\gamma \equiv \frac{1-\kappa}{3+\widetilde{\epsilon}+\varrho\max(\|\bar{x}\|^2/3+\|\epsilon\|_\infty, 1)}$, then we have

$$x_k \to x^\star \in \operatorname{Argmin}(f) \cap B\left(\overline{\mathcal{X}}, \varsigma\right)$$

and $\exists K > 0$ such that for all $k \geq K$, we have

$$\operatorname{dist}^2(x_k, \overline{\mathcal{X}}) \leq (1-\nu)^{k-K}\rho^2 + \varsigma^2, \tag{1.3.2}$$

this holds with high probability where $\nu \in [0, 1[$.

(ii) Suppose that ϱ obeys (3.6.12). If m is such that $m \geq C(\varrho, \|\epsilon\|_\infty)n\log(n)$, and Algorithm 3 is initialized with the spectral method in Algorithm 4, then (1.3.2) holds for all $k \geq K = 0$ with high probability.
The rate $1-\nu$ is given explicitly in Theorem 4.3.2.

The reconstruction error is eventually ς and this will be shown to scale as $O\left(\frac{\|\epsilon\|}{\sqrt{m}\|\bar{x}\|}\right)$ which is minimax optimal according to [63, Theorem 3].

Though we focus on Gaussian measurements when establishing the global recovery properties of our mirror descent algorithm, our theory extends to the situation where the a_r's are i.i.d sub-Gaussian random vectors.

For the CDP model, we only have guarantees in absence of noise. For instance, we show that one can afford a smaller sampling complexity bound but at the price of using an appropriate spectral initialization procedure to find an initial guess near a solution before applying our scheme. Starting from this initial guess, mirror descent then converges linearly to the true vector up to a global sign change with a dimension-independent convergence rate. Our main result for the CDP measurements model is the following.

Theorem 1.3.3 (Exact recovery guarantee for CDP model).
Let $\varrho \in]0,1[$ and $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1.

(i) If the number of patterns P satisfies $P \geq C(\varrho)\log(n)$, then with high probability, for almost all initializers x_0 of Algorithm 1 used with constant step-size $\gamma_k \equiv \gamma = \frac{1-\kappa}{L}$, for any $\kappa \in]0,1[$ and L given by Lemma 3.2.3, $(x_k)_{k\in\mathbb{N}}$ converges to a critical point which cannot be a strict saddle.

(ii) Let $\delta \in]0, \min(\|\bar{x}\|^2, 1)/2[$. There exists $\rho_\delta > 0$ such that if ϱ is small enough and $P \geq C(\varrho)n\log^3(n)$, and if Algorithm 1 is initialized with the spectral method in Algorithm 2, then with high probability

$$\operatorname{dist}^2(x_k, \overline{\mathcal{X}}) \leq (1-\nu)^k, \quad \forall k \geq 0, \tag{1.3.3}$$

where $\nu \in [0,1[$ and will be given explicitly.

The CDP model is very challenging. Indeed, at this stage, we do not have any theoretical guarantee with the CDP model, neither for generic strict saddle points avoidance nor for stable recovery by mirror descent. One of the main difficulties is that several of our arguments rely on uniform bounds, for instance on the Hessian, that need to hold simultaneously for all vectors $x \in \mathbb{R}^n$ with high probability. But the CDP model bears much less randomness to exploit for establishing such bounds with reasonable sampling complexity bounds. Nevertheless, numerical experiments suggest that global convergence and stable recovery still holds for our mirror descent algorithm with CDP measurements. A rigorous analysis of these numerical results is open and is left for future research.

1.3.2 Regularized phase retrieval

Our contributions are along two lines. First, we propose and analyze an inertial Bregman proximal gradient algorithm to solve ($\mathcal{P}$) and then we turn to studying noiseless and stable phase retrieval with general low complexity promoting regularizers.

1.3.2.1 Inertial Bregman proximal gradient under partial smoothness

This part is purely algorithmic. We study the global and local convergence properties of an inertial type Bregman proximal gradient (see Algorithm 5) to solve the problem ($\mathcal{P}$) with a class of Bregman kernels that satisfies the triangle scaling property (see Definition 2.3.9). Our main motivation is that the Tikhonov variational formulation ($\mathcal{P}_{y,\lambda}$) of the regularized phase retrieval satisfies the list of Assumptions 5.1.1-5.2.1. Besides, $F_{y,\lambda}$ is semialgebraic. We can sum up the main contributions of this work in the following compact theorem.

Theorem 1.3.4 (Convergence analysis of the IBPG). *Let us consider the problem ($\mathcal{P}$) with Assumption 5.1.1-5.2.1 and assume that Φ is a semialgebraic function. Besides, let us assume that the inertial parameters $(a_k)_{k\in\mathbb{N}}$ converge to $a \in [0,1]$. Suppose that the sequence $(z_k)_{k\in\mathbb{N}}$ generated by Algorithm 5 is bounded, then*

(i) *all the sequences* $(x_k)_{k\in\mathbb{N}}, (y_k)_{k\in\mathbb{N}}$ *and* $(z_k)_{k\in\mathbb{N}}$ *have finite length and converges to a critical point.*

(ii) *if the algorithm is started near a global minimizer* $x^\star$ *in the* $\Phi-$*attentive topology, the generated sequences converge to* $x^\star$*.*

(iii) *if* $x_\star \in \mathrm{crit}\Phi$ *is the limit of the sequence and assume that* G *is partly smooth at* $x_\star$ *relative to a manifold* $\mathcal{M}_{x_\star}$ *and that a non-degeneracy condition holds at* $x_\star$*. Then there exists a constant* K *large enough such that for all* $k \geq K, x_k \in \mathcal{M}_{x_\star}$*.*

(iv) *if* F *is locally* C^2 *around the cluster point* $x_\star$ *where now both a non-degeneracy and a restricted injectivity condition hold, then* $(x_k)_{k\in\mathbb{N}}$ *converges locally linearly to* $x_\star$*.*

(v) *If* $G \equiv 0$*, then for almost all initializers, the generated sequences converge toward the set of critical points that are not strict saddle points.*

This result shows that for semialgebraic (and more generally tame) functions, bounded iterates generated by the inertial Bregman proximal gradient converge to an element in the set of critical points of the objective function. If we start near an optimal solution the sequence converges to it. Besides, this result shows that the inertial Bregman proximal gradient enjoys a finite activity identification property. In the case, we carry out a sharp spectral analysis from which we exhibit a linear convergence regime under a particular choice of inertial parameters. This choice of parameters highly depends on the triangle scaling exponent κ, which generalizes the Euclidean case when this parameter is just 2. When $G \equiv 0$, for almost all initializers the inertial Bregman proximal gradient avoids the strict saddle points, and we recall that these are critical points where the function has a direction of negative curvature.

1.3.2.2 Low-complexity regularized phase retrieval

Noiseless recovery We establish sufficient conditions under which solving the minimization problem ($\mathcal{P}_{\bar{y},0}$) recovers the original signal $\bar{x}$ up to a global sign change. These conditions are deterministic and depend on the regularizer R (for instance its descent cone at $\bar{x}$) and the measurement matrix A. This holds true with high probability for standard Gaussian measurements and common regularizers, given a sufficient number of measurements. Furthermore, we derive precise recovery guarantees for specific regularizer types, including decomposable ones (like Lasso and group Lasso), frame analysis-based regularizers, and the total variation. The following theorem highlights our deterministic results.

Theorem 1.3.5 (Exact recovery). *Assume that the set of minimizers of* ($\mathcal{P}_{\bar{y},0}$) *is nonempty. If* R *is a proper convex lower semicontinuous and even function obeying* (**H.2**)*, then the recovery of* $\bar{x}$ *(up to a global sign) is exact by solving* ($\mathcal{P}_{\bar{y},0}$)*.*

Stable recovery We analyze the minimization problem ($\mathcal{P}_{y,\lambda}$) which is a noise-aware with a Tikhonov-type regularization. We show that under appropriate nondegeneracy and restricted injectivity conditions, the solution of this problem converges to the original signal (up to sign change) as $\lambda \to 0$ when $\epsilon \to 0$. We also give robust recovery error bounds for small enough noise. For standard Gaussian measurements and various regularizers, we provide sample complexity bounds for the above conditions to hold, and in turn, for robust recovery to occur.

Theorem 1.3.6 (Stable recovery). *Consider the noisy phaseless measurements in* (GeneralPR)*. Let* $\sigma \overset{\text{def}}{=} \|\epsilon\|$*. Assume that* R *is a nonnegative convex lower semicontinuous and even function obeying* (**H.2**)*. Suppose also that* R *is coercive on* $\ker(A)$*, and that*

$$\lambda \to 0 \ \text{ and } \ \sigma^2/\lambda \to 0, \quad \text{as} \quad \sigma \to 0.$$

Then, for any minimizer $x^\star_{y,\lambda}$ *of* ($\mathcal{P}_{y,\lambda}$)

$$\operatorname{dist}\left(x^\star_{y,\lambda}, \overline{\mathcal{X}}\right) \to 0 \quad as \quad \sigma \to 0.$$

If R *and* A *verify an appropriate nondegeneracy and restricted injectivity conditions (see* (6.4.4) *and* (6.4.5)*) at* $\bar{x}$*, then for* σ *small enough and* $\lambda \propto \sigma$

$$\operatorname{dist}\left(x^\star_{y,\lambda}, \overline{\mathcal{X}}\right) = O(\sigma).$$

1.4 Outline

The remainder of the thesis is divided into two parts and seven chapters.

Chapter 2: Mathematical Background. This chapter provides the necessary mathematical material used throughout the manuscript. It gathers tools from convex analysis, as well as Bregman based optimization, tools from nonsmooth analysis, and elements from probability theory such as concentration inequalities.

Chapter 3: Phase retrieval using mirror descent. We start by describing the mirror descent algorithm with backtracking and establish its global and local convergence guarantees in the deterministic case applied to phase retrieval. We then turn to the case of random measurements and we provide sample complexity bounds for the deterministic guarantees to hold with high probability. The last section is devoted to the numerical experiments.

Chapter 4: Stable Phase retrieval using mirror descent. We first establish convergence guarantees of mirror descent in the deterministic case for noisy phase retrieval. We then instantiate to the case of Gaussian measurements and provide sample complexity bounds for the deterministic guarantees to hold with high probability. Numerical experiments close this chapter.

Chapter 5: Inertial Bregman proximal gradient under Partial smoothness. We start with the global convergence analysis of the inertial Bregman proximal gradient scheme under the Kurdyka-Łojasiewicz property. We then delve into a local analysis under partial smoothness. Trap avoidance in the smooth case is also investigated. Numerical experiments on regularized phase retrieval are then described.

Chapter 6: Low-complexity regularized Phase retrieval. We consider the regularized phase retrieval. We establish perfect recovery of the true signal up to global sign change from the minimization problem in the noiseless setting. We then turn to the noisy case, where we analyze two formulations of the problem. Finally, we study convergence of the minimizers to the set of true vectors.

Chapter 7: Conclusion & Perspectives. This last chapter sums up the main ideas and contributions of the thesis and draws important conclusions. It also discusses several interesting perspectives and open problems that we believe are worth investigating in the future.

1.5 Work Not Included in the Thesis

To close the introduction, I list here works that are not included in this thesis. I give short summaries of them outlining the main ideas without delving into the details. These works were carried out with collaborators from the Concept team at the Fresnel Institute throughout the PhD. They were intended to prepare the application of our phase retrieval framework and algorithms to light scattering and precision optics.

1.5.1 Immediate and one-point roughness measurements using spectrally shaped light.

This work appeared in [45]. We propose a novel approach different from the usual scattering measurements, one that is free of any mechanical movement or scanning. Scattering is measured along a single direction. Wide-band illumination with a properly chosen wavelength spectrum makes the signal proportional to the sample roughness, or to the higher-order roughness moments. Spectral shaping is carried out with gratings and a spatial light modulator. We validate the technique by crosschecking with a classical angle- resolved scattering set-up. Though the bandwidth is reduced, this white light technique may be of key interest for on-line measurements, large components that cannot be displaced, or other parts that do not allow mechanical movement around them.

1.5.2 Instantaneous measurement of surface roughness spectra using white light scattering projected on a spectrometer.

This work was published in [46]. We propose a new experiment of white light scattering that should overtake the previous ones in most situations. The set-up is very simple, as it requires only a broadband illumination source and a spectrometer to analyze light scattering at a unique direction. After introducing the principle of the instrument, roughness spectra are extracted for different samples and the consistency of results is validated at the intersection of bandwidths. The technique will be of great use for samples that cannot be moved.

Chapter 2

Background

Contents

We assemble in this chapter the relevant background material and some notations that will be used throughout the following chapters.

2.1 Notations

Vectors and matrices We denote $\langle \cdot, \cdot \rangle$ the scalar product on $\mathbb{R}^n$ and $\|\cdot\|$ the corresponding norm. $B(x, r)$ is the corresponding ball of radius r centered at x and $\mathbb{S}^{n-1}$ is the corresponding unit sphere. Moreover, $\|\cdot\|_p, p \in [1, \infty]$ stands for the ℓ_p norm . For $m \in \mathbb{N}^*$, we use the shorthand notation $[\![m]\!] = \{1, \dots, m\}$. The i-th entry of a vector x is denoted $x[i]$. For any $y \in \mathbb{R}^m$ the operations $|y|$ and y^2 should be understood componentwise. Given a matrix $M \in \mathbb{R}^{m \times n}$, $M^\top$ is its transpose. Let $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ be respectively the smallest and the largest eigenvalues of M. For two real symmetric matrices M and N, $M \succeq N$ if $M - N$ is positive semidefinite. For a linear operator M, M^* is its adjoint.

In the following, or a subspace $V \subset \mathbb{R}^n$, P_V denotes the orthogonal projector on V, and

$$x_V = \mathrm{P}_V x \ \text{ and } \ A_V = A\mathrm{P}_V.$$

For $I \subset [\![m]\!]$, $A^I \stackrel{\text{def}}{=} [a_i : \ i \in I]^\top$ denotes the sub-matrix whose rows are only those of A indexed by I. We denote $|I|$ the cardinality of I and I^c its complement.

Throughout, we use the shorthand notation $\overline{\mathcal{X}} \stackrel{\text{def}}{=} \{\pm \bar{x}\}$ to denote the set of true vectors. Hence, for any vector $x \in \mathbb{R}^n$, the distance to the set of true vectors is

$$\operatorname{dist}(x, \overline{\mathcal{X}}) \stackrel{\text{def}}{=} \min(\|x - \bar{x}\|, \|x + \bar{x}\|). \tag{2.1.1}$$

Remark 2.1.1. Our limitation of the set of true solutions to $\{\pm \bar{x}\}$ may appear restrictive since even for real vectors, the equivalence class is much larger than what we are allowing. Moreover, our restriction will be justified in the oversampling regime with random measurements. For instance,

for Gaussian measurements, only $\{\pm\bar{x}\}$ are provably global minimizers for large enough number of measurements. Moreover, for the two types of random measurements on which we focus in this thesis, spectral initialization also provides an initialization which is real and provably lies in the neighborhood of $\{\pm\bar{x}\}$.

2.2 Nonsmooth and Convex Analysis

Sets For a nonempty set $\mathcal{S} \in \mathbb{R}^n$, we denote $\overline{\mathcal{S}}$ it closure, $\overline{\text{conv}}\,(\mathcal{S})$ the closure of its convex hull, and $\iota_{\mathcal{S}}$ its *indicator function i.e.*,

$$\iota_{\mathcal{S}}(x) \stackrel{\text{def}}{=} \begin{cases} 0 & x \in \mathcal{S} \\ +\infty & x \notin \mathcal{S}. \end{cases}$$

Recall that, if $\mathcal{S}$ is nonempty, closed, and convex, then $\iota_{\mathcal{S}}$ belongs to $\Gamma_0\,(\mathbb{R}^n)$. For a nonempty convex set $\mathcal{S}$, its *affine hull* $\text{aff}(\mathcal{S})$ is the smallest affine manifold containing it. It is a translate of its *parallel subspace* $\text{par}(\mathcal{S})$, i.e. $\text{par}(\mathcal{S}) = \text{aff}(\mathcal{S}) - \mathcal{S} = \mathbb{R}(\mathcal{S} - \mathcal{S})$, for any $x \in \mathcal{S}$. The *relative interior* $\text{ri}(\mathcal{S})$ of a convex set $\mathcal{S}$ is the interior of $\mathcal{S}$ for the topology relative to its affine full.

For any vector $x \in \mathbb{R}^n$, the distance to a non-empty set $\mathcal{S} \subset \mathbb{R}^n$ is

$$\text{dist}(x, \mathcal{S}) \stackrel{\text{def}}{=} \inf_{z \in \mathcal{S}} \|x - z\|\,. \tag{2.2.1}$$

We recall that the orthogonal projection of $x \in \mathbb{R}^n$ on $\mathcal{S}$ is define by

$$\text{P}_{\mathcal{S}}(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{z \in \mathcal{S}} \|x - z\|\,. \tag{2.2.2}$$

Definition 2.2.1 (Support function). The *support function* of $\mathcal{S} \subset \mathbb{R}^n$ is

$$\sigma_{\mathcal{S}}(z) = \sup_{x \in \mathcal{S}} \langle z, x \rangle\,.$$

Definition 2.2.2 (Polar set). Let $\mathcal{S}$ be a nonempty convex set. The set $\mathcal{S}^\circ$ given by

$$\mathcal{S}^\circ = \{v \in \mathbb{R}^n :\ \langle v, x \rangle \leq 1 \quad \forall x \in \mathcal{S}\}$$

is called the *polar of* $\mathcal{S}$.

The set $\mathcal{S}^\circ$ is closed convex and contains the origin. When $\mathcal{S}$ is also closed and contains the origin, then it coincides with its bipolar, i.e. $\mathcal{S}^{\circ\circ} = \mathcal{S}$.

Definition 2.2.3 (Gauge). Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a non-empty closed convex set containing the origin. The *gauge* of $\mathcal{S}$ is the function $\gamma_{\mathcal{S}}$ defined on $\mathbb{R}^n$ by

$$\gamma_{\mathcal{S}}(x) = \inf\{\lambda > 0 :\ x \in \lambda\mathcal{S}\}.$$

As usual, $\gamma_{\mathcal{S}}(x) = +\infty$ if the infimum is not attained.

We have the following characterization of the support function in finite dimension. $\gamma_{\mathcal{S}}$ is a nonnegative, closed and sublinear function. When $\mathcal{S}$ is a closed convex set containing the origin, then

$$\gamma_{\mathcal{S}} = \sigma_{\mathcal{S}^\circ} \text{ and } \gamma_{\mathcal{S}^\circ} = \sigma_{\mathcal{S}}.$$

Let $\mathcal{S} \subset \mathbb{R}^n$ a nonempty, closed bounded and convex subset. If $0 \in \text{ri}(\mathcal{S})$, then $\gamma_{\mathcal{S}} \in \Gamma_0(\mathbb{R}^n)$ is sublinear, nonnegative and finite-valued, and

$$\sigma_{\mathcal{C}}(x) = 0 \iff x \in (\text{par}(\mathcal{C}))^\perp.$$

Definition 2.2.4 (Asymptotic cone). Let $\mathcal{S}$ be a non-empty closed convex set. The *asymptotic cone*, or *recession cone* is the closed convex cone defined by

$$\mathcal{S}_\infty \stackrel{\text{def}}{=} \bigcap_{t>0} \frac{\mathcal{S} - x}{t}, \quad x \in \mathcal{S}.$$

This definition does not depend on the choice of $x \in \mathcal{S}$. The importance of the asymptotic cone becomes obvious through the following fundamental fact; see [12, Proposition 2.1.2].

Fact 2.2.5. $\mathcal{S}$ is compact if and only if $\mathcal{S}_\infty = \{0\}$.

Functions

A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed (or lower semicontinuous (lsc)) if its epigraph is closed. It is coercive if

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

The effective domain of f is $\mathrm{dom}(f) = \{x \in \mathbb{R}^n : \ f(x) < +\infty\}$ and f is proper if $\mathrm{dom}(f) \neq \emptyset$ as is the case when it is finite-valued.

A function is said sublinear if it is convex and positively homogeneous. The Legendre-Fenchel conjugate of f is

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

Let the kernel of a function be defined as $\ker(f) \stackrel{\text{def}}{=} \{z \in \mathbb{R}^n : f(z) = 0\}$. Let us denote by $\mathcal{S}_{\mathrm{lev}f}(\bar{x}) \stackrel{\text{def}}{=} \{z \in \mathbb{R}^n : f(z) \leq f(\bar{x})\}$ the sublevel set of f at $\bar{x}$. We denote by $\Gamma_0(\mathbb{R}^n)$ the class of proper lsc convex function from $\mathbb{R}^n$ to $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$. For $\mu > 0$, f is μ-strongly convex if and only if $f - \frac{\mu}{2} \|\cdot\|^2$ is convex, and is μ-weakly convex (or semiconvex) if and only if $f + \frac{\mu}{2} \|\cdot\|^2$ is convex.

A point $x \in \mathbb{R}^n$ is in the $f-$attentive neighborhood of $x_\star \in \mathbb{R}^n$, if for all $r > 0$, there exists $\rho \in]0, r[$ and $\eta > 0$ such that $\|x - x_\star\| \leq \rho$ and $f(x_\star) < f(x) < f(x_\star) + \eta$. The following notation $x \xrightarrow{f} x_\star$ stand for $f-$attentive convergence, *i.e.*, $x \to x_\star$ with $f(x) \to f(x_\star)$.

Definition 2.2.6. (Subdifferentials) The Fréchet subdifferential of f at a point $x \in \mathbb{R}^n$ is the set

$$\partial_F f(x) \stackrel{\text{def}}{=} \begin{cases} \{v \in \mathbb{R}^n : \ f(z) \geq f(x) - \langle v; z - x \rangle + o(\|z - x\|)\}, & \text{if } x \in \mathrm{dom}(f) \\ \emptyset, \quad \text{otherwise.} \end{cases}$$

The limiting subdifferential at x is defined as the set

$$\partial f(x) \stackrel{\text{def}}{=} \{v \in \mathbb{R}^n : \ \exists x_k \xrightarrow{f} x, v \leftarrow v_k \in \partial_F f(x_k)\}.$$

An element of $\partial f(x)$ is called a subgradient. If f is differentiable at x, then its only subgradient is its gradient, *i.e.*, $\partial_F f(x) = \partial f(x) = \{\nabla f(x)\}$. While $\partial_F f$ is convex-valued, ∂f is closed-valued. If f is (subdifferentially) regular at x then both subdifferentials coincide. This is the case in particular for any $f \in \Gamma_0(\mathbb{R}^n)$, and in this case ∂f is the usual (Fenchel) subdifferential in the sense of convex analysis

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

The set of critical points of f is $\mathrm{crit}(f) = \{x_\star \in \mathbb{R}^n : \ 0 \in \partial f(x_\star)\}$, and $\mathrm{Argmin}(f)$ is the set of global minimizers of f.

Definition 2.2.7 (Asymptotic function). For a proper closed function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $f_\infty : \mathbb{R}^n \to \overline{\mathbb{R}}$ is the *asymptotic function*, or *recession function* associated with f, which is defined by

$$f_\infty(z) \stackrel{\text{def}}{=} \liminf_{z' \to z, t \to +\infty} \frac{f(tz')}{t}. \tag{2.2.3}$$

It is well-known that f_∞ is lsc and positively homogeneous and that its epigraph is the asymptotic cone of the epigraph of f. This function plays an important role in the existence of solutions to minimization problems. Besides for any closed convex set $\mathcal{S}$, one has

$$(\iota_{\mathcal{S}})_\infty = \iota_{\mathcal{S}_\infty}.$$

The following result relates coercivity to properties of the recession function.

Proposition 2.2.8. *Let $f \in \Gamma_0(\mathbb{R}^n)$ and $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator. Then,*

(i) *g coercive* $\iff$ $f^\infty(x) > 0 \quad \forall x \neq 0$.

(ii) $f^\infty \equiv \sigma_{\mathrm{dom}(f^*)}$.

(iii) $(f \circ A)^\infty \equiv f^\infty \circ A$.

In particular, we deduce that $f \circ A$ is coercive if and only if $\sigma_{\mathrm{dom}(f^*)}(Ax) > 0$ for every $x \neq 0$.

Proof. The proofs can be found in [156, Theorem 3.26], [156, Theorem 11.5] and [105, Corollary 3.2] respectively. □

Operator norm Let g_1 and g_2 be two finite-valued gauges defined on two vector spaces V_1, V_2, and $A : V_1 \to V_2$ be a linear map. The *operator bound* $\|A\|_{g_1 \to g_2}$ of A between g_1 and g_2 is given by

$$\|A\|_{g_1 \to g_2} = \sup_{g_1(x) \leq 1} g_2(Ax).$$

Let us note that $\|A\|_{g_1 \to g_2} < \infty$ if and only if $A\ker(g_1) \subset \ker(g_2)$. Moreover a sufficient condition for $\|A\|_{g_1 \to g_2} < \infty$ is that g_1 is coercive. As a convention, $\|A\|_{g_1 \to \|\cdot\|_p}$ is denoted as $\|A\|_{g_1 \to \ell^p}$. A direct consequence of this definition is the fact that, for every $x \in V_1$,

$$g_2(Ax) \leq \|A\|_{g_1 \to g_2} g_1(x).$$

2.3 Bregman Toolbox

Definition and properties Let us start with the definition of a Legendre function.

Definition 2.3.1. [155, Chapter 26]**(Legendre function)** Let $\phi \in \Gamma_0(\mathbb{R}^n)$ such that $\mathrm{int}(\mathrm{dom}(\phi)) \neq \emptyset$. ϕ is called

(i) *essentially smooth* if it is differentiable on $\mathrm{int}(\mathrm{dom}(\phi))$ with $\|\nabla\phi(x_k)\| \to \infty$ for every sequence $(x_k)_{k\in\mathbb{N}}$ of $\mathrm{int}(\mathrm{dom}(\phi))$ converging to a boundary point of $\mathrm{dom}(\phi)$.

(ii) *essentially strictly convex* if it is strictly convex on every convex subset of $\mathrm{dom}\,\partial\phi \overset{\text{def}}{=} \{x : \partial\phi(x) \neq \emptyset\}$.

A Legendre function is essentially smooth and strictly convex.

Remark 2.3.2. [155, Theorem 26.5]

- Let us notice that a function is Legendre if and only if its conjugate ϕ^* is of Legendre.
- We also have that $\mathrm{dom}\partial\phi = \mathrm{int}(\mathrm{dom}(\phi))$, $\partial\phi = \emptyset, \forall x \in \mathrm{bd}(\mathrm{dom}(\phi))$ and $\forall x \in \mathrm{int}(\mathrm{dom}(\phi))$ we have $\partial\phi(x) = \{\nabla\phi(x)\}$ and $\nabla\phi$ is a bijection from $\mathrm{int}(\mathrm{dom}(\phi))$ to $\mathrm{int}(\mathrm{dom}(\phi))^*$ with $\nabla\phi^* = (\nabla\phi)^{-1}$.

For any function $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$, we define a proximity measure associated with ϕ.

Definition 2.3.3. (Bregman divergence) The general Bregman divergence associated with ϕ is

$$D_\phi^v(x,y) \overset{\text{def}}{=} \begin{cases} \phi(x) - \phi(y) - \langle v; x - y\rangle, & \text{if} (x,y) \in (\mathrm{dom}(\phi) \times \mathrm{int}(\mathrm{dom}(\phi))), v \in \partial\phi(y), \\ +\infty & \text{otherwise.} \end{cases} \tag{2.3.1}$$

Remark 2.3.4. When ϕ is Legendre or simply sufficiently smooth on $\mathrm{int}(\mathrm{dom}(\phi))$, we recover the classical definition *i.e.*,

$$D_\phi(x,y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle. \tag{2.3.2}$$

If $\phi(x) = \frac{1}{2}\|x\|^2$, the Bregman divergence is the usual euclidean distance $D_\phi(x,y) = \frac{1}{2}\|x - y\|^2$. This proximity measure is not a distance (it's not symmetric in general for instance).

Throughout the rest of the work, we use the following properties of the Bregman divergence.

Proposition 2.3.5. ***(Properties of the Bregman distance)***

(i) D_ϕ *is nonnegative if and only if* ϕ *is convex. If in addition,* ϕ *is strictly convex,* D_ϕ *vanishes if and only if its arguments are equal.*

(ii) Linear additivity: for any $\alpha, \beta \in \mathbb{R}$ *and any functions* ϕ_1 *and* ϕ_2 *sufficiently smooth, we have*

$$D_{\alpha\phi_1+\beta\phi_2}(x,u) = \alpha D_{\phi_1}(x,u) + \beta D_{\phi_2}(x,u), \tag{2.3.3}$$

for all $(x,u) \in (\mathrm{dom}\phi_1 \cap \mathrm{dom}\phi_2)^2$ *such that both* ϕ_1 *and* ϕ_2 *are differentiable at* u*.*

(iii) The three-point identity: For any $x \in \mathrm{dom}(\phi)$ *and* $u, z \in \mathrm{int}(\mathrm{dom}(\phi))$*, we have*

$$D_\phi(x,z) - D_\phi(x,u) - D_\phi(u,z) = \langle \nabla\phi(u) - \nabla\phi(z); x-u\rangle\,. \tag{2.3.4}$$

(iv) Suppose that ϕ *is also* $C^2(\mathrm{int}(\mathrm{dom}(\phi)))$ *and* $\nabla^2\phi(x)$ *is positive definite for any* $x \in \mathrm{int}(\mathrm{dom}(\phi))$*. Then for every convex compact subset* $\Omega \subset \mathrm{int}(\mathrm{dom}(\phi))$*, there exists* $0 < \theta_\Omega \leq \Theta_\Omega < +\infty$ *such that for all* $x, u \in \Omega$*,*

$$\frac{\theta_\Omega}{2}\|x-u\|^2 \leq D_\phi(x,u) \leq \frac{\Theta_\Omega}{2}\|x-u\|^2\,. \tag{2.3.5}$$

Regularity of functions The following definition extends the classical gradient Lipschitz continuity property to the Bregman setting, this notion is named "relative smoothness" and is important to the analysis of optimization problems that are differentiable but lack of gradient Lipschitz-smoothness. The earliest reference to this notion can be found in an economics paper [34] where it is used to address a problem in game theory involving fisher markets. Later on it was developed in [21, 40] and then in [124], although first coined relative smoothness in [124]. Let $\phi \in \Gamma_0(\mathbb{R}^n) \cap C^1(\mathrm{int}(\mathrm{dom}(\phi)))$, and g be a proper and lower semicontinuous function such that $\mathrm{dom}(\phi) \subset \mathrm{dom}(g)$.

Definition 2.3.6. **($L-$relative smoothness)** Let $g \in C^1(\mathrm{int}(\mathrm{dom}(\phi)))$, g is called $L-$smooth relative to ϕ on $\mathrm{int}(\mathrm{dom}(\phi))$ if there exists $L > 0$ such that $L\phi - g$ is convex on $\mathrm{int}(\mathrm{dom}(\phi))$, *i.e.*

$$D_g(x,u) \leq L D_\phi(x,u) \quad \text{for all} \quad (x,u) \in \mathrm{dom}(\phi) \times \mathrm{int}(\mathrm{dom}(\phi)). \tag{2.3.6}$$

When ϕ is the energy entropy, *i.e.* $\phi = \frac{1}{2}\|\cdot\|^2$, one recovers the standard descent lemma implied by Lipschitz continuity of the gradient of g.

In a similar way, we also extend the standard local strong convexity property to a relative version *w.r.t* to an entropy or kernel ϕ.

Definition 2.3.7. **(Local relative strong convexity)** Let $\mathcal{C}$ be a non-empty subset of $\mathrm{dom}(\phi)$. Let $g \in C^1(\mathrm{int}(\mathrm{dom}(\phi)))$, for $\sigma > 0$ we say that g is σ-strongly convex on $\mathcal{C}$ relative to ϕ if

$$D_g(x,u) \geq \sigma D_\phi(x,u) \quad \text{for all} \quad x \in \mathcal{C}\,\text{and}\,u \in \mathcal{C} \cap \mathrm{int}(\mathrm{dom}(\phi)). \tag{2.3.7}$$

When $\mathcal{C} = \mathrm{dom}(\phi)$, we get the idea of global relative strong convexity. If ϕ is the energy entropy (*i.e.* $\phi = \frac{1}{2}\|\cdot\|^2$), one recovers the standard definition of (local/global) strong convexity.

The idea of global (*i.e.* $\mathcal{C} = \mathrm{dom}(\phi)$) relative strong convexity has already been used in the literature, see *e.g.* [169, Proposition 4.1] and [20, Definition 3.3]. Its local version was first proposed in [163]. When ϕ is the energy entropy (*i.e.* $\phi = \frac{1}{2}\|\cdot\|^2$), one recovers the standard definition of (local/global) strong convexity. Relation of global relative strong convexity to gradient dominated inequalities, which is an essential ingredient to prove global linear convergence of mirror descent, was studied in [20, Lemma 3.3].

Let us give the following useful lemma which compare the Bregman divergences of smooth functions.

Lemma 2.3.8. *Let* $g, \phi \in C^2(\mathbb{R}^n)$*. If* $\forall u \in \mathbb{R}^n$*,* $\nabla^2 g(u) \preceq \nabla^2\phi(u)$ *for all* u *in the segment* $[x,z]$*, then,*

$$D_g(x,z) \leq D_\phi(x,z). \tag{2.3.8}$$

Proof. The result comes from the Taylor-MacLaurin expansion. Indeed we have $\forall x, z \in \mathbb{R}^n$

$$\begin{aligned} D_g(x,z) &= g(x) - g(z) - \langle \nabla g(z), x - z \rangle \\ &= \int_0^1 (1-\tau) \left\langle x - z, \nabla^2 g(z + \tau(x-z))(x-z) \right\rangle d\tau, \end{aligned}$$

and thus

$$D_\phi(x,z) - D_g(x,z) = \int_0^1 (1-\tau) \left\langle x - z, \left(\nabla^2 \phi(z + \tau(x-z)) - \nabla^2 g(z + \tau(x-z))\right)(x-z) \right\rangle d\tau.$$

The positive semidefiniteness assumption implies the claim. □

Triangle scaling property Here, we introduce the triangle scaling property (TSP) [89] for Bregman distances.

Definition 2.3.9. Let ϕ be a Legendre function. The Bregman distance generated by ϕ has the triangle scaling property if there is a constant $\kappa > 0$ such that for all $x, y, z \in \mathrm{ri}(\mathrm{dom}(\phi))$,

$$D_\phi((1-a)x + ay, (1-a)x + az) \leq a^\kappa D_\phi(y,z), \quad \forall a \in [0,1]. \tag{2.3.9}$$

We call κ the uniform triangle scaling exponent (TSE) of D_ϕ.

There is a large class of functions that satisfy this property, here are some specific examples.

- *Euclidean distance.* When ϕ is the energy and thus $D_\phi(x,y) = \frac{1}{2}\|x-y\|^2$. The squared Euclidean distance has a uniform TSE $\kappa = 2$.
- *Bregman divergence induced by strongly convex and smooth functions.* If ϕ is $\sigma_\phi-$strongly convex and $L-$smooth over its domain then (2.3.9) hold with $\kappa = 2$ if the right-hand side is multiplied by the condition number L/σ_ϕ.
- *Bregman geometry based on polynomial kernel.* Polynomial functions of the form $\phi(x) = \frac{1}{p}\|x\|^p$ for some $p \geq 2$, the global TSE for the induced Bregman divergence can be less than 1 for $p > 2$. However, the modified reference function $\phi(x) = \frac{1}{2}\|x\|^2 + \frac{1}{p}\|x\|^p$ for $p \geq 4$ has a coefficient $\kappa > 1$, or $\kappa = 2$ with an additional factor on the right-hand side of (2.3.9), over a bounded domain. It turns that this choice of ϕ is precisely the one that we make for phase retrieval as announced in (1.2.2).

We have the following proposition.

Proposition 2.3.10. ***(Second order characterization of the TSP.)*** *Let $\phi \in C^2(\mathbb{R}^n)$ Legendre function. If the kernel or entropy function satisfies the TSP then we have for all $u \in [(1-a)x + ay, (1-a)x + az]$ and for all $v \in [y,z]$ then*

$$a^\kappa \nabla^2 \phi(v) - a^2 \nabla^2 \phi(u) \succeq 0. \tag{TSP}$$

Proof. The proof follows from Taylor expansion of C^2-smooth functions. □

2.4 KL Functions

This section encompasses all the essential components required for the axiomatization of convergence for KL functions and therefore can be skipped by an experienced reader. We start by defining the non-smooth KL property which is an additional assumption on the class of functions that we consider. This property gives a hint about the geometric bearing of the function near the point where it is satisfied.

Definition 2.4.1. (Non-smooth KL property) A proper and lower semicontinuous function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ has the KL property at a point $x_\star \in \mathrm{dom}(g)$ if there exists a neighborhood $U_{x_\star}$, $\eta > 0$ and a concave real-valued function $\varphi \in C^1([0, \eta[)$, with $\varphi(0) = 0$ and $\varphi' > 0$, such that

$$\varphi'\left(g(x) - g(x_\star)\right) \mathrm{dist}(0, \partial g(x)) \geq 1, \quad \forall x \in U_{x_\star} \cap \left\{x \in \mathbb{R}^n : \ g(x_\star) < g(x) < g(x_\star) + \eta\right\}.$$

If g has the KL property at each point of $\mathrm{dom}(g)$, g is called a KL function.

φ is known as the desingularizing function. The KL property is also closely related to error bounds and the broader notion of "(sub)metric regularity". We refer the reader to [96] and [37] for a detailed study of these notions. In general, it is not obvious to check whether a given function is KL or not. Actually, this is a very deep question that has been studied at the interface of analysis and algebraic geometry. For smooth functions, it has been shown that semi-algebraic and sub-analytic are Łojaciewicz in the seminal works of [121, 122, 104]. This has been extended to the nonsmooth case and then widely studied in the scope of optimization in [36, 35, 37]. For instance, functions definable on o-minimal structures are KL. This cover most functions studied in practice, and for instance those in this manuscript.

The following uniformization of the KL property will be very useful; see [38, Lemma 6].

Lemma 2.4.2 (Uniformized KL property). *Let Ω be a compact set, and let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper and closed function. Assume that g is constant on Ω and satisfies the KL property at each point of Ω. Then, there exist $\varepsilon > 0$, $\eta > 0$, if there exists $\eta > 0$ and a concave real-valued function $\varphi \in C^1([0, \eta[)$ with $\varphi(0) = 0, \varphi' > 0$ such that for all $x_\star$ in Ω, one has*

$$\varphi'(g(x) - g(x_\star))\mathrm{dist}(0, \partial g(x)) \geq 1,$$

and all $x \in \mathbb{R}^n$ such that $\mathrm{dist}(x, \Omega) < \varepsilon$ and $g(x_\star) < g(x) < g(x_\star) + \eta$.

For the convergence of our inertial Bregman proximal gradient (IBPG) algorithm, we will use a general convergence mechanism as first axiomatized in [8] for descent algorithms and generalized in [38] on the so-called PALM algorithm, so that it can be used and applied to any given algorithm such as ours (see also, *e.g.* [40, Appendix 6] for a self-contained presentation). The main goal is to prove that the whole sequence $(x_k)_{k\in\mathbb{N}}$ generated by IBPG, converges to a critical point. For that purpose, considering a Lyapunov function Ψ associated to IBPG, it has to satisfy the following three key conditions.

Definition 2.4.3 (Descent-like method). A sequence $(x_k)_{k\in\mathbb{N}}$ is called descent-like for the function Ψ if the following conditions hold:

(C.1) *Sufficient decrease condition.* There exists a positive scalar ρ_1 such that

$$\rho_1 \left\|x_k - x_{k-1}\right\|^2 \leq \Psi(x_k, x_{k-1}) - \Psi(x_{k+1}, x_k), \quad \forall k \in \mathbb{N}.$$

(C.2) *Relative error condition.* There exists $K \in \mathbb{N}$ and $\rho_2 > 0$ such that $\forall k \geq K$, there exists $v_{k+1} \in \partial\Psi(x_{k+1}, x_k)$ such that

$$\|v_{k+1}\| \leq \rho_2 \left(\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|\right).$$

(C.3) *Continuity condition.* Let x^* be a limit point of a subsequence $(x_k)_{k\in\mathcal{K}\subset\mathbb{N}}$ then we have that $\limsup_{k\in\mathcal{K}\subset\mathbb{N}} \Psi(x_k, x_{k-1}) \leq \Psi(x^*) \overset{\text{def}}{=} \Psi(x^*, x^*)$.

Condition (C.1) is intended to model a descent property of the Lyapunov function, and hence a dissipation of the energy Ψ. (C.2)[1] originates from the well-known fact that most algorithms in

[1]The original version of this condition in [8] involves only the first term in the bound. The reasoning however remains the same with this version of the inequality; see *e.g.* [41, 117, 136].

optimization generate sequences via exact or inexact minimization of subproblems and condition (C.2) reflects relative inexact optimality conditions for such minimization subproblems. Condition (C.3) is a weak requirement which, in particular, holds when Ψ is continuous. However, the latter is not mandatory in general as the nature of the algorithm (IBPG here) will force the sequences to comply with (C.3) under a simple lower semicontinuity assumption.

Equipped with Definition 2.4.3, and when Ψ satisfies the KL property, the following global convergence result holds true.

Theorem 2.4.4 (Global convergence). *Let $(x_k)_{k\in\mathbb{N}}$ be a bounded sequence generated by a descent-like method for Ψ. If Ψ satisfies the KL property, then the sequence $(x_k)_{k\in\mathbb{N}}$ has finite length, i.e., $\sum_{k\in\mathbb{N}} \|x_{k+1} - x_k\| < +\infty$ and it converges to $x_\star \in \mathrm{crit}(\Psi)$.*

2.5 Riemannian Geometry and Partial Smoothness

2.5.1 Riemannian geometry

In this section, we introduce the essential tools from Riemannian geometry that appear throughout this work. This is done to make the manuscript self-contained and its reading smoother.

Let $\mathcal{M}$ be a C^2-smooth embedded submanifold of $\mathbb{R}^n$ around a point $x \in \mathbb{R}^n$. With some abuse of terminology, we will state C^2-manifold instead of C^2-smooth embedded submanifold of $\mathbb{R}^n$. The natural embedding of a submanifold $\mathcal{M}$ into $\mathbb{R}^n$ permits to define a Riemannian structure and to introduce geodesics on $\mathcal{M}$, and we simply say $\mathcal{M}$ is a Riemannian manifold. We denote respectively $\mathcal{T}_\mathcal{M}(x)$ and $\mathcal{N}_\mathcal{M}(x)$ the tangent and normal space of $\mathcal{M}$ at $x \in \mathcal{M}$.

Exponential map Geodesics generalize the concept of straight lines from linear spaces to manifolds. It is a smooth curve from an interval of $\mathbb{R}$ to $\mathcal{M}$, with intrinsic acceleration normal everywhere to $\mathcal{M}$. Roughly speaking, it is locally the shortest path between two points on $\mathcal{M}$. Let denote by $\mathfrak{g}(t;x,h)$ the value at $t \in \mathbb{R}$ of the geodesic starting $\mathfrak{g}(0;x,h) = x \in \mathcal{M}$ with velocity $\dot{\mathfrak{g}}(t;x,h) = \dfrac{d\mathfrak{g}(t;x,h)}{dt} = h \in \mathcal{T}_\mathcal{M}(x)$ (uniquely defined). It is important to realize that for every $h \in \mathcal{T}_\mathcal{M}(x)$ there exists an interval I around 0 and a unique geodesic $\mathfrak{g}(t;x,h) : I \to \mathcal{M}$ such that $\mathfrak{g}(0;x,h) = x$ and $\dot{\mathfrak{g}}(0;x,h) = t$. The mapping $\mathrm{Exp}_x : \mathcal{T}_\mathcal{M}(x) \to \mathcal{M}, \quad h \mapsto \mathrm{Exp}_x(h) = \mathfrak{g}(1;x,h)$, is called the *Exponential map*. Given $x, x' \in \mathcal{M}$, and a direction $h \in \mathcal{T}_\mathcal{M}(x)$ we are want such map to fulfill $\mathrm{Exp}_x(h) = x' = \mathfrak{g}(1;x,h)$.

Parallel translation Given two points $x, x' \in \mathcal{M}$ let $\mathcal{T}_\mathcal{M}(x), \mathcal{T}_\mathcal{M}(x')$ be the corresponding tangent spaces. Define $\tau : \mathcal{T}_\mathcal{M}(x) \to \mathcal{T}_\mathcal{M}(x')$, the *parallel translation* along the unique geodesic joining x to x', which is an isomorphism and isometry with respect to the Riemannian metric.

Riemannian gradient and Hessian For a vector $v \in \mathcal{N}_\mathcal{M}(x)$, the Weingarten map of $\mathcal{M}$ at x is the operator $\mathfrak{W}_x(.,v) : \mathcal{T}_\mathcal{M}(x) \to \mathcal{T}_\mathcal{M}(x)$ defined by:

$$\mathfrak{W}_x(h,v) = -\mathrm{P}_{\mathcal{T}_\mathcal{M}(x)} dV[h],$$

where V is any local extension of v to a normal vector field on $\mathcal{M}$. The definition does not depend of the choice of the extension V. The Weingarten map as defined above is a symmetric linear operator which is closely tied to the second fundamental form of $\mathcal{M}$ ([43, Definition 5.48]). Let g be a real-valued function which is C^2 along $\mathcal{M}$ around x. The covariant gradient of g at $x' \in \mathcal{M}$ is the vector denoted $\nabla_\mathcal{M} g(x') \in \mathcal{T}_\mathcal{M}(x')$ defined by:

$$\langle \nabla_\mathcal{M} g(x'), h \rangle = \frac{d}{dt} g(\mathrm{P}_\mathcal{M}(x' + th))|_{t=0}, \quad \forall h \in \mathcal{T}_\mathcal{M}(x'),$$

where $\mathrm{P}_{\mathcal{M}}$ is the projection onto $\mathcal{M}$. The covariant Hessian of g at x' is the symmetric linear mapping $\nabla^2_{\mathcal{M}} g(x')$ from $\mathcal{T}_{\mathcal{M}}(x')$ to itself which is defined as

$$\left\langle \nabla^2_{\mathcal{M}} g(x')h; h \right\rangle = \frac{d^2}{dt^2} g(\mathrm{P}_{\mathcal{M}}(x'+th))|_{t=0}, \quad \forall h \in \mathcal{T}_{\mathcal{M}}(x').$$

This definition agrees with the definition using geodesics or connections. Now, assume that $\mathcal{M}$ is a Riemannian embedded submanifold of $\mathbb{R}^n$ and g has a C^2-smooth restriction on $\mathcal{M}$. This can be characterization by the existence of a C^2- smooth extension (representative) of g *i.e.* a C^2-smooth function $\widetilde{g}$ on $\mathbb{R}^n$ such that $\widetilde{g}$ agrees with g on $\mathcal{M}$. Thus, the Riemannian gradient $\nabla_{\mathcal{M}} g(x')$ is also given by

$$\nabla_{\mathcal{M}} g(x') = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla \widetilde{g}(x'),$$

and $\forall h \in \mathcal{T}_{\mathcal{M}}(x')$, the Riemannian Hessian reads

$$\begin{aligned} \nabla^2_{\mathcal{M}} g(x')h &= \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} d\left(\nabla_{\mathcal{M}} g(x')\right)[h] = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} d\left(\nabla_{\mathcal{M}} \widetilde{g}(x')\right)[h] \\ &= \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla^2 \widetilde{g}(x')h + \mathfrak{W}_{x'}(h, \mathrm{P}_{\mathcal{N}_{\mathcal{M}}(x')} \nabla \widetilde{g}(x')). \end{aligned}$$

When $\mathcal{M}$ is affine or linear subspace of $\mathbb{R}^n$, then obviously $\mathcal{M} = x + \mathcal{T}_{\mathcal{M}}(x)$ and $\mathfrak{W}_{x'}(h, \mathrm{P}_{\mathcal{N}_{\mathcal{M}}(x')} \nabla \widetilde{g}(x')) = 0$ and finally

$$\nabla^2_{\mathcal{M}} g(x') = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla^2 \widetilde{g}(x') \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')}.$$

The next two lemmas will be instrumental when analyzing the local convergence behaviour of our inertial algorithm. We refer to [115, Section 2.6] for their proofs.

Lemma 2.5.1. *Let $x \in \mathcal{M}$ and x_k a sequence converging to x in $\mathcal{M}$. Denote $\tau_k : \mathcal{T}_{\mathcal{M}}(x_k) \to \mathcal{T}_{\mathcal{M}}(x_k)$ be the parallel translation along the unique geodesic joining x to x_k. Then, for any bounded vector $u \in \mathbb{R}^n$, we have:*

$$\left(\frac{1}{\tau_k} \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x_k)} - \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)} \right) u = o\left(\|u\|\right). \tag{2.5.1}$$

Lemma 2.5.2. *Let x, x' be two close points in $\mathcal{M}$, denote $\tau : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(x')$ the parallel translation along the unique geodesic joining x too x'. The Riemannian Taylor expansion of $g \in C^2(\mathcal{M})$ around x reads,*

$$\frac{1}{\tau} \nabla_{\mathcal{M}} g(x') = \nabla_{\mathcal{M}} g(x) + \nabla^2_{\mathcal{M}} g(x) \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)}(x' - x) + o(\|x' - x\|). \tag{2.5.2}$$

2.5.2 Partial Smoothness

Introduced by Lewis in [111], "Partial smoothness" captures the characteristics of the geometry of nonsmooth functions. It axiomatizes the notion of active/identifiable submanifold or identifiable surfaces in [186]. A partly smooth function is smooth along the identifiable submanifold and sharp transversally to the manifold. Therefore, the behaviour of the function of its minimizers depends essentially on its restriction to this manifold, hence offering a powerful framework for algorithmic and sensitivity analysis theory.

Definition 2.5.3 (Partly smooth function). A function $g \in \Gamma_0(\mathbb{R}^n)$ is C^2-partly smooth at a point x relative to the set $\mathcal{M}$ containing x, if $\partial g(x) \neq \emptyset$ and $\mathcal{M}$ is an embedded C^2-smooth submanifold and there exists a neighborhood $\mathcal{V}_x$ of x such that the following properties hold

(i) **(Smoothness)** the restriction $g_{|\mathcal{M}}$ is a C^2 function in the neighborhood $\mathcal{V}_x$;

(ii) **(Sharpness)** The affine hull of $\partial g(x)$ is a translation of the space $\mathcal{N}_{\mathcal{M}}(x)$, *i.e.*

$$S_x \overset{\text{def}}{=} \mathrm{par}(\partial g(x)) = \mathcal{N}_{\mathcal{M}}(x) \Leftrightarrow T_x \overset{\text{def}}{=} \mathrm{par}(\partial g(x))^{\perp} = \mathcal{T}_{\mathcal{M}}(x).$$

(iii) **(Continuity)** The set-valued mapping ∂g is continuous at x relative to $\mathcal{M}$.

Observe that $S_x = T_x^{\perp}$ by definition. Throughout the rest of the work, we denote the class of C^2-partly smooth function at x relative to $\mathcal{M}$ by $\mathrm{PSF}_x(\mathcal{M})$.

Owing to the definition of partial smoothness, we have the following facts.

Fact 2.5.4. (Local normal sharpness) If $g \in \mathrm{PSF}_x(\mathcal{M})$, then for all point $x' \in \mathcal{M}$ near x we have $\mathcal{T}_{\mathcal{M}}(x') = T_{x'}$. In particular when $\mathcal{M}$ is affine or linear, then $T_{x'} = T_x$.

Fact 2.5.5. If $g \in \mathrm{PSF}_x(\mathcal{M})$, then for all $x' \in \mathcal{M}$ near x we have

$$\nabla_{\mathcal{M}} g(x') = \mathrm{P}_{T_{x'}}\left(\partial g(x')\right),$$

and this does not depend on the smooth representation of g on $\mathcal{M}$. In turn, for all $h \in T_{x'}$,

$$\nabla^2_{\mathcal{M}} g(x')h = \mathrm{P}_{T_{x'}} \nabla^2 \widetilde{g}(x')h + \mathfrak{W}_{x'}(h, \mathrm{P}_{T_{x'}^{\perp}} \nabla \widetilde{g}(x')),$$

where $\widetilde{g}$ is a smooth representative of g on $\mathcal{M}$ and $\mathfrak{W}_{x'}(\cdot,\cdot)$ is the Weingarten map of $\mathcal{M}$ at x.

2.6 Probability and Concentration Inequalities

Many of the following notations for probabilistic concepts are adopted directly from [176, 171]. We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ a *probability space* with a *set of events* Ω, a *σ-algebra* $\mathcal{F}$, and a *probability measure* $\mathbb{P}$.

Definition 2.6.1. Let $\mathcal{S}$ be an arbitrary bounded subset of $\mathbb{R}^n$. The covering number of $\mathcal{S}$ in the Euclidean norm at resolution $\delta > 0$ is the smallest number, $N(\mathcal{S}, \delta)$, such that $\mathcal{S}$ can be covered with balls $B(x_i, \delta)$, $x_i \in \mathcal{S}$, $i \in [\![N(\mathcal{S}, \delta)]\!]$, i.e.,

$$\mathcal{S} \subseteq \bigcup_{i \in [\![N(\mathcal{S},\delta)]\!]} B(x_i, \delta)$$

The finite set of points $\mathcal{S}_\delta \overset{\text{def}}{=} \{x_i : i \in [\![N(\mathcal{S}, \delta)]\!]\}$ is called a δ-covering or δ-net of $\mathcal{S}$.

Definition 2.6.2. The Gaussian width of a subset $\mathcal{S} \subset \mathbb{R}^n$ is defined as

$$w(\mathcal{S}) \overset{\text{def}}{=} \mathbb{E}\left(\sigma_{\mathcal{S}}(g)\right), \quad \text{where} \quad g \sim \mathcal{N}(0, \mathrm{Id}_n).$$

The Gaussian width is a summary geometric quantity that, informally speaking, measures the size of the bulk of a set in $\mathbb{R}^n$. This concept plays a central role in high-dimensional probability and its applications. It has appeared in the literature in different contexts [85]. In particular, it has been used to establish sample complexity bounds to ensure exact recovery (noiseless case) and mean-square estimation stability (noisy case) for low-complexity penalized estimators from Gaussian measurements; see e.g. [61, 5, 145, 173]. The Gaussian width has deep connections to convex geometry and it enjoys many useful properties. It is well-known that it is positively homogeneous, monotonic w.r.t inclusion, and invariant under orthogonal transformations. Moreover, one has

$$w(\mathcal{S}) = w(\overline{\mathcal{S}}) = w(\mathrm{conv}\,(\mathcal{S})) = w(\overline{\mathrm{conv}}\,(\mathcal{S})).$$

This comes from the properties of the support function. A lower bound for the Gaussian width of a bounded set can be obtained via Sudakov's minoration.

Proposition 2.6.3. *Let $\mathcal{S}$ be a bounded set. Then for any $\delta > 0$ small enough, we have*

$$w(\mathcal{S}) \geq \delta \sqrt{\log\left(N(\mathcal{S}, \delta)\right)}.$$

Proof. Let $\mathcal{S}_\delta$ be an δ-net of $\mathcal{S}$. Thus,

$$w(\mathcal{S}) \geq w(\mathcal{S}_\delta).$$

Since $\min_{x_i \neq x_j \subset \mathcal{S}_\delta} \|x_i - x_j\| = 2\delta$, the claim follows from [42, Theorem 13.4] for all δ smaller than the diameter of $\mathcal{S}_\delta$. □

Next, we recall some deviation and concentration inequalities that will be important for us.

Proposition 2.6.4 (Markov inequality). *Let X be a random variable and φ a nondecreasing nonnegative function then $\forall t > 0$ such that $\varphi(t) > 0$ we have*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(\varphi(X))}{\varphi(t)}.$$

Proposition 2.6.5 (Tchebychev inequality). *Let X be a random variable with finite variance σ^2. Then $\forall t > 0$ we have*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t\sigma) \leq \frac{1}{t^2}.$$

For a random variable X and $k \geq 1$, we define

$$\|X\|_{\psi_k} = \sup_{p \geq 1} p^{-1/k} (\mathbb{E}(|X|^p))^{1/p}.$$

$\|X\|_{\psi_2}$ is known as the sub-Gaussian norm while $\|X\|_{\psi_1}$ is the sub-exponential norm.

Proposition 2.6.6 (Hoeffding-type inequality). *Let $X = (X_1, \cdots, X_N)$ be independent centered sub-gaussian random variables, and let $K = \max_i \|X_i\|_{\psi_2}$. Then for every vector $a \in \mathbb{R}^N$ and $t \geq 0$, we have*

$$\mathbb{P}(|\langle a, X \rangle| \geq t) \leq e. \exp\left(-\frac{ct^2}{K^2 \|a\|^2}\right),$$

where $c > 0$ is an absolute constant.

Proposition 2.6.7 (Bernstein-type inequality). *Let $X_1, \cdots, X_N$ be independent centered sub-exponential random variables, and let $K = \max_i \|X_i\|_{\psi_1}$. Then for every vector $a \in \mathbb{R}^N$ and $t \geq 0$, we have*

$$\mathbb{P}(|\langle a, X \rangle| \geq t) \leq e. \exp\left\{-c \min\left(\frac{t^2}{K^2 \|a\|^2}, \frac{t}{K \|a\|_\infty}\right)\right\},$$

where $c > 0$ is an absolute constant.

The following proposition gives the concentration of measure in the Gauss space. A comprehensive account can be found in [108].

Proposition 2.6.8. *Let f be a real-valued $K-$Lipschitz continuous on $\mathbb{R}^n$. Let g be the standard normal random vector in $\mathbb{R}^n$. Then for every $t \geq 0$ one has*

$$\Pr\{f(g) - \mathbb{E}(f(g)) \geq t\} \leq \exp(-t^2/2K^2).$$

Part I

Phase Retrieval without Regularization

Chapter 3

Provable Phase Retrieval with Mirror Descent

In this chapter, we consider the problem of phase retrieval, which consists of recovering an n-dimensional real vector from the magnitude of its m linear measurements. We propose a mirror descent (or Bregman gradient descent) algorithm based on a wisely chosen Bregman divergence, hence allowing to remove the classical global Lipschitz continuity requirement on the gradient of the non-convex phase retrieval objective to be minimized. We apply the mirror descent for two random measurements: the i.i.d. standard Gaussian and those obtained by multiple structured illuminations through Coded Diffraction Patterns (CDP). For the Gaussian case, we show that when the number of measurements m is large enough, then with high probability, for almost all initializers, the algorithm recovers the original vector up to a global sign change. For both measurements, the mirror descent exhibits a local linear convergence behaviour with a dimension-independent convergence rate. Our theoretical results are finally illustrated with various numerical experiments, including an application to the reconstruction of images in precision optics. Our main contributions and findings can be summarized as follows:

Main contributions of this chapter

- For general sensing vectors, bounded iterates of our algorithm converge to a critical point which is not a strict saddle point. In addition, provided that a local relative strong convexity, the mirror descent exhibits a local linear convergence behaviour.
- For Gaussian standard measurements, when the number of sensing vector is large enough for almost all initializer the mirror descent recovers the true signal up to a global sign change with a local linear convergence which is dimension-independent.
- For CDP and Gaussian measurements, if we afford a smaller sampling complexity we have to use an appropriate initialization method to be close the true signal. Then starting from this initial guess, mirror descent converges linearly to the true vector up to a global sign change with a dimension-independent convergence rate.

The content of this chapter appeared in [83].

Contents

3.1 Introduction

3.1.1 Problem Statement

In this chapter, we consider the noiseless version of phase retrieval problem (GeneralPR), *i.e.*, $\epsilon = 0$, that we recall for convenience. Let $\bar{x} \in \mathbb{R}^n$ be a vector to be recovered and that we are given information about the squared modulus of the inner product between $\bar{x}$ and m sensing/measurement vectors $(a_r)_{r\in[\![m]\!]}$. The noiseless phase retrieval problem can be cast as:

$$\begin{cases} \text{Recover } \bar{x} \in \mathbb{R}^n \text{ from the measurements } y \in \mathbb{R}^m \\ y[r] = |a_r^* \bar{x}|^2, \quad r \in [\![m]\!], \end{cases} \tag{NLPR}$$

where $[r]$ is the r-th entry of the corresponding vector. Throughout the chapter, A is the $m \times n$ matrix with a_r^*'s as its rows.

Since $\bar{x}$ is real-valued, the best one can hope is to ensure that $\bar{x}$ is uniquely determined by y up to a global sign. Phase retrieval is in fact an ill-posed inverse problem in general and is known to be NP-hard [157]. Thus, one of the major challenges is to design efficient recovery algorithms and find conditions on m and $(a_r)_{r\in[\![m]\!]}$ which guarantee exact recovery (up to a global sign change); see Section 1.2.1 for a review and discussion of the state-of-the-art.

3.1.2 Contributions and relation to prior work

In this chapter, we cast (NLPR) as solving the minimization problem (3.2.1). Inspired by [40], we propose a mirror descent (or Bregman gradient descent) algorithm with backtracking associated to a wisely chosen Bregman divergence, hence removing the classical global Lipschitz continuity requirement on the gradient of the nonconvex objective in (3.2.1).

In the deterministic case, we show that for almost all initializers, bounded iterates of our algorithm converge to a critical point where the objective has no direction of negative curvature, i.e., a critical point which is not a strict saddle point. In addition, provided that a local relative strong convexity property holds, we also show that our mirror descent scheme exhibits a local linear convergence behaviour.

In the case of i.i.d standard Gaussian measurements, provided that the the number m of sensing vectors is large enough, it turns out that the iterates of our algorithm are bounded, and that the set of critical points of the objective f in (3.2.1) is the union of $\{\pm\bar{x}\}$ and the set of strict saddle points. This together with the above deterministic guarantees ensures that with high probability, for almost all initializers, our mirror descent recovers the original vector $\bar{x}$ up to a global sign change, and exhibits a local linear convergence behaviour with a dimension-independent convergence rate. Our results are far more general than those of [64] as we require for instance a smaller sampling complexity bound and we assume any random initialization provided that it is drawn from a distribution that has a density *w.r.t* the Lebesgue measure, *i.e.* the Gaussian nature of initialization in [64] is irrelevant in our context.

For both CDP and Gaussian measurements, we show that one can afford a smaller sampling complexity bound but at the price of using an appropriate spectral initialization procedure to find an initial guess near a solution before applying our scheme. Starting from this initial guess, mirror descent then converges linearly to the true vector up to a global sign change with a dimension-independent convergence rate. This is in contrast with the Wirtinger flow [53] which also requires spectral initialization and whose local convergence rate degrades with the dimension, though the latter aspect has been improved in the truncated Wirtinger flow [63]. The Polyak subgradient method [69] initialized with a spectral method provably converges linearly with isotropic sub-gaussian measurements under a sample complexity bound similar to ours. However, no analysis is known for the CDP measurement model. Observe also that the Polyak subgradient algorithm requires the knowledge of the minimal value of the phase retrieval objective. This is obviously 0 for the noiseless case but is unknown in the noisy one. In terms of computational complexity, mirror descent involves solving the mirror step (see Proposition 3.2.4) which amounts to computing the unique real positive root of a third order polynomial and then multiplying it by the entry vector. This costs $O(n)$ operations. Overall, the computational complexity of mirror descent is similar to that of other first-order methods such as the Wirtinger flow or the Polyak subgradient algorithm.

Though we focus on Gaussian measurements when establishing the global recovery properties of our mirror descent algorithm, our theory extends to the situation where the a_r's are i.i.d sub-Gaussian random vectors. The case where a_r's are a drawn form the CDP model is, however, far more challenging. One of the main difficulties is that several of our arguments rely on uniform bounds, for instance on the Hessian, that need to hold simultaneously for all vectors $x \in \mathbb{R}^n$ with high probability. But the CDP model bears much less randomness to exploit for establishing such bounds with reasonable sampling complexity bounds. Whether this is possible or not is an open problem that we leave to future research.

3.1.3 Chapter organization

The rest of the chapter is organized as follows. In Section 3.2, we describe the mirror descent algorithm with backtracking and establish its global and local convergence guarantees in the deterministic case. We then turn to the case of random measurements in Section 3.3 where we provide sample complexity bounds for the deterministic guarantees to hold with high probability. Section 3.4 is devoted to the numerical experiments. The proofs of technical results are collected in Section 3.5 and Section 3.6.

3.2 Deterministic Phase Retrieval

3.2.1 Phase retrieval minimization problem

In this work, we cast (NLPR) as solving the following optimization problem

$$\min_{x\in\mathbb{R}^n}\left\{f(x)\overset{\text{def}}{=}\frac{1}{4m}\sum_{r=1}^{m}\left(y[r]-|(Ax)[r]|^2\right)^2\right\}. \tag{3.2.1}$$

Observe that $f\in C^2(\mathbb{R}^n)$ but is obviously nonconvex. Actually, f is weakly convex (or semiconvex).

Proposition 3.2.1. *f is μ-weakly convex with $\mu=m^{-1}\sum_{r=1}^m|y[r]|\,\|a_r\|^2$.*

Proof. Starting from the Hessian of f in (3.5.2) and using Cauchy-Schwarz inequality, we have for any $z\in\mathbb{R}^n$,

$$m\left\langle\nabla^2 f(x)z,z\right\rangle=3\sum_{r=1}^{m}|(Ax)[r]|^2|(Az)[r]|^2-\sum_{r=1}^{m}y[r]|(Az)[r]|^2\geq-\sum_{r=1}^{m}|y[r]|\,\|a_r\|^2\,\|z\|^2.$$

Recalling that f is μ-weakly convex if and only if $\nabla^2 f(x)+\mu\mathrm{Id}\succeq 0$, we conclude. □

It is also clear that ∇f is not Lipschitz continuous. This is the main motivation behind considering the framework of Bregman gradient descent. As we will see shortly, f has a relative smoothness property (see Definition 2.3.6 above) with respect to a well-chosen entropy function. In turn, relative smoothness will prove crucial for establishing descent properties of Bregman gradient descent, also known as, mirror descent.

Following [40], let us consider the following kernel or entropy function

$$\psi(x)=\frac{1}{4}\|x\|^4+\frac{1}{2}\|x\|^2. \tag{3.2.2}$$

Proposition 3.2.2. *ψ enjoys the following properties:*

(i) $\psi\in C^2(\mathbb{R}^n)$, is 1-strongly convex and Legendre according to Definition 2.3.1.

(ii) $\nabla\psi$ is Lipschitz over bounded subsets of $\mathbb{R}^n$.

(iii) $\nabla\psi$ is a bijection from $\mathbb{R}^n$ to $\mathbb{R}^n$, and its inverse is $\nabla\psi^*$.

The first two claims are easy to show. The last one follows from [155, Theorem 26.5].

It turns out that the objective f in (3.2.1) is smooth relative to the entropy ψ defined in (3.2.2) on the whole space $\mathbb{R}^n$. This is stated in the following result whose proof is provided in Section 3.5.1.

Lemma 3.2.3. *Let f and ψ as defined in (3.2.1) and (3.2.2) respectively. f is L-smooth relative to ψ on $\mathbb{R}^n$ for any $L\geq\frac{1}{m}\sum_{r=1}^m 3\,\|a_r\|^4$.*

This estimate of of the modulus of relative smoothness L in Lemma 3.2.3 is rather crude but has the advantage to not depend on the measurements y. A far sharper estimate will be provided in the case where the sensing vectors are random; see Section 3.3.

3.2.2 Mirror descent with backtracking

We recall the following mapping closely related to the Bregman gradient descent. For all $x \in \mathbb{R}^n$ and any step-size $\gamma > 0$,

$$T_\gamma(x) \overset{\text{def}}{=} \operatorname*{argmin}_{u \in \mathbb{R}^n} \left\{ \langle \nabla f(x), u - x \rangle + \frac{1}{\gamma} D_\psi(u, x) \right\}. \tag{3.2.3}$$

The pair (f, ψ) defined in (3.2.1)-(3.2.2) satisfies [40, Assumptions A, B, C, D] (in fact ψ is even strongly convex in our case). Therefore, it is straightforward to see that T_γ is a well-defined and single-valued on $\mathbb{R}^n$; see [40, Lemma 3.1]. Moreover, by virtue of Proposition 3.2.2, letting $x^+ = T_\gamma(x)$, the first order optimality condition for (3.2.3) reads

$$x^+ = F(x) \overset{\text{def}}{=} \nabla\psi^{-1}\left(\nabla\psi(x) - \gamma\nabla f(x)\right) = \nabla\psi^*\left(\nabla\psi(x) - \gamma\nabla f(x)\right). \tag{3.2.4}$$

Our mirror descent (or Bregman gradient descent) scheme with backtracking is summarized in Algorithm 1.

Algorithm 1: Mirror Descent for Phase Retrieval

Parameters: $0 < L_0 \leq L$ (see Lemma 4.2.1), $\kappa \in]0, 1[$, $\xi \leq 1$.
Initialization: $x_0 \in \mathbb{R}^n$;
for $k = 0, 1, \ldots$ **do**
  repeat
    $\gamma_k = \frac{1-\kappa}{L_k}$;
    $x_{k+1} = F(x_k) = \nabla\psi^*\left(\nabla\psi(x_k) - \gamma_k \nabla f(x_k)\right)$;
    $L_k \leftarrow L_k/\xi$;
  until $D_f(x_{k+1}, x_k) \leq \xi L_k D_\psi(x_{k+1}, x_k)$;
  $L_{k+1} \leftarrow \xi L_k$;
  Output: x_{k+1}.

Observe that Algorithm 1 cannot be trapped in the second loop thanks to Lemma 4.2.1. Indeed, we have $L_k \in [L_0, L/\xi]$ for all $k \in \mathbb{N}$. The version without backtracking is recovered by setting $\xi = 1$ and using constant step-size verifying $\gamma \in]0, 1/L[$ where L is the global relative smoothness coefficient. Backtracking for an inertial version of the Bregman proximal gradient algorithm was used in [136].

It remains now to compute the mirror step. This amounts to finding a root of a third-order polynomial.

Proposition 3.2.4. ***(Mirror step computation)*** *[40, Proposition 5.1] Let $x \in \mathbb{R}^n$ and $p_\gamma(x) = \nabla\psi(x) - \gamma\nabla f(x)$. Then computing (3.2.4) amounts to*

$$x^+ = t^* p_\gamma(x), \tag{3.2.5}$$

where t^* is the unique real positive root of $t^3 \left\|p_\gamma(x)\right\|^2 + t - 1 = 0$.

3.2.3 Deterministic recovery guarantees by mirror descent

We pause to recall two notions that will be important in our convergence result.

Definition 3.2.5. (f-attentive neighborhood) A point $u \in \mathbb{R}^n$ belongs to an f-attentive neighborhood of $x \in \mathbb{R}^n$, if there exist $\delta > 0$ and $\mu > 0$ such that $u \in B(x, \delta)$ and $f(x) < f(u) < f(x) + \mu$.

Definition 3.2.6. (Strict saddle points) A point $x_\star \in \operatorname{crit}(f)$ is a strict saddle point of f if $\lambda_{\min}(\nabla^2 f(x_\star)) < 0$. The set of strict saddle points of f is denoted $\operatorname{strisad}(f)$.

We are now ready to state our main convergence result.

Theorem 3.2.7. *Let $(x_k)_{k\in\mathbb{N}}$ be a bounded sequence generated by Algorithm 1 for the phase retrieval problem (NLPR). Then,*

(i) the sequence $(f(x_k))_{k\in\mathbb{N}}$ is non-increasing,

(ii) the sequence $(x_k)_{k\in\mathbb{N}}$ has a finite length and converges to a point in $\mathrm{crit}(f)$.

(iii) Let $r > 0$. Assume that the initial point x_0 is in the f-attentive neighborhood of $x^\star \in \mathrm{Argmin}(f) \neq \emptyset$, i.e. $\exists \delta \in]0, r[$ and $\mu > 0$ such that $x_0 \in B(x^\star, \delta)$ and $f(x_0) \in]0, \mu[$, then

(a) $\forall k \in \mathbb{N}, x_k \in B(x^\star, r)$, and x_k converges to a global minimizer of f.

(b) Besides, if $\exists \rho > 0$ such that f is σ-strongly convex on $B(x^\star, \rho)$ relative to ψ, with $r \leq \frac{\rho}{\max(\sqrt{\Theta(\rho)},1)}$, where we recall $\Theta(\rho)$ from Proposition 2.3.5-(iv), then $\forall k \in \mathbb{N}$

$$\|x_k - x^\star\|^2 \leq \left(\prod_{i=0}^{k-1} \frac{1-\sigma\gamma_i}{1+\sigma\gamma_i\Theta(\rho)^{-1}}\right)\rho^2 \to 0. \tag{3.2.6}$$

(iv) If $L_k = L$, then for Lebesgue almost all initializers x_0, the sequence $(x_k)_{k\in\mathbb{N}}$ converges to an element in $\mathrm{crit}(f)\backslash\mathrm{strisad}(f)$.

See Section 3.5.2 for the proof.

Remark 3.2.8.

- A standard assumption that automatically guarantees the boundedness of the sequence $(x_k)_{k\in\mathbb{N}}$, hence its convergence to a critical point, is coercivity of f. Since the latter is a composition of a coercive function (a positive quartic function) and the linear operator A (recall that its rows are the a_r^*'s), coercivity of f amounts to injectivity of A. This is exactly what we will show in the random case when m is large enough.
- It is clear that $\mathrm{Argmin}(f) \neq \emptyset$ since $\overline{\mathcal{X}} \subset \mathrm{Argmin}(f)$ and the claim (iii) applies at $\pm\bar{x}$ in which case one has exact recovery up to a global sign.
- A close inspection at the proof of Proposition 2.3.5-(iv) shows that $\Theta(\rho) = \sup_{x\in B(\bar{x},\rho)} \|\nabla^2\psi(x)\|$ does the job. In view of (3.5.3), it is easy to see that $\Theta(\rho) \leq 6\|\bar{x}\|^2 + 6\rho^2 + 1$.
- Claim (iii) shows local linear convergence of x_k to $x^\star$. Indeed, $\sigma \leq L_k$ for any k, and thus $1 - \sigma\gamma_k \in]\kappa, 1[$.
- Clearly, claim (iv) states that when the initial point is selected according to a distribution which has a density *w.r.t* the Lebesgue measure, then the sequence $(x_k)_{k\in\mathbb{N}}$ converges to a point that avoids strict saddle points of f. This is a consequence of the centre stable manifold theorem applied to our mirror descent algorithm.
- When it will come to the phase retrieval problem from random measurements (see forthcoming section), in order to prove local linear convergence, the key argument will be to show that for a sufficient number of measurements, then *w.h.p* f is strongly convex around $\pm\bar{x}$ relative to ψ.

3.3 Random Phase Retrieval via Mirror Descent

3.3.1 Framework

Throughout the chapter, we will work under two random measurement models:

(1) The sensing vectors are drawn i.i.d following a (real) standard Gaussian distribution. We can then rewrite the observation data as

$$y[r] = |a_r^\top \bar{x}|^2, \quad r \in [\![m]\!], \tag{3.3.1}$$

where $(a_r)_{r\in[\![m]\!]}$ are i.i.d $\mathcal{N}(0,1)$.

(2) The Coded Diffraction Patterns (CDP) model, as considered for instance in [52]. The idea is to modulate the signal before diffraction in the case of the Fourier transform measurements. The observation model is then

$$y = \left(|\mathcal{F}(D_p\bar{x})[j]|^2\right)_{j,p} = \left(\left|\sum_{\ell=0}^{n-1} \bar{x}_\ell d_p[\ell] e^{-i\frac{2\pi j\ell}{n}}\right|^2\right)_{j,p}. \tag{3.3.2}$$

where $j \in \{0,\dots,n-1\}$ and $p \in \{0,\dots,P-1\}$, D_p is a real diagonal matrix with the modulation pattern d_p on its diagonal, and $\mathcal{F}$ is the discrete Fourier transform. P is the number of coded patterns/masks and the total number of measurements is then $m = nP$. The modulation patterns $(d_p)_{p\in[P]}$ are i.i.d copies of the same random vector d satisfying the following assumption:

Assumption 3.3.1.

(A.1) *d is symmetric and $\exists M > 0$ such that $|d| \le M$.*

(A.2) *Moments conditions: $\mathbb{E}(d) = 0$ and $\mathbb{E}(d^4) = 2\mathbb{E}(d^2)^2$. Without loss of generality, we assume $\mathbb{E}(d^2) = 1$.*

For example, we can take ternary random variables with values in $\{-1,0,1\}$ with probabilities $\{\frac{1}{4},\frac{1}{2},\frac{1}{4}\}$. We refer to [52] for other modulation patterns.

When the number of measurements is large enough for both measurements models, we will be able to establish local convergence properties of Algorithm 1 provided it is initialized with a good guess. For this, we use a spectral initialization method; see for instance [53, 63, 141, 192, 182, 179]. The procedure consists of taking x_0 as the leading eigenvector of a specific matrix as described in Algorithm 2.

Algorithm 2: Spectral Initialization.

Input: $y[r], r = 1,\dots,m$.

Output: x_0

Set $\lambda^2 = n\frac{\sum_r y[r]}{\sum_r \|a_r\|^2}$;

Take x_0 the top eigenvector of $Y = \frac{1}{m}\sum_{r=1}^m y[r]a_r a_r^*$ normalized to $\|x_0\| = \lambda$.

Remark 3.3.2. Assuming random measurements models and using probabilistic arguments to get sample complexity bounds and understand fundamental limits of phase retrieval (and other inverse problems) is an established technique in the applied mathematics literature. Of course, we are aware that this might not always be realistic from an application perspective as it may sometimes involve changing the data measurements to fit the theory. Nonetheless, for the application we have in mind (precision in optics), the CDP measurement model seems reasonable. This is the subject of an ongoing work.

We are now ready to state our main results for each measurement model.

3.3.2 Gaussian measurements

Before stating our result, we consider the following events which will be helpful in our proofs. For this, we fix $\varrho \in]0,1[$ and $\lambda \in]0,1[$.

- The event

$$\mathcal{E}_{\text{strictsad}} = \left\{\text{crit}(f) = \overline{\mathcal{X}} \cup \text{strisad}(f)\right\} \tag{3.3.3}$$

means that the set of critical points of the function f is reduced to $\{\pm\bar{x}\}$ and the set of strict saddle points.

- The event
$$\mathcal{E}_{\text{conH}} = \left\{\forall x \in \mathbb{R}^n, \quad \left\|\nabla^2 f(x) - \mathbb{E}\left(\nabla^2 f(x)\right)\right\| \leq \varrho\left(\|x\|^2 + \|\bar{x}\|^2/3\right)\right\} \tag{3.3.4}$$
captures the deviation of the Hessian of f around its expectation.
- The event
$$\mathcal{E}_{\text{inj}} = \left\{\forall x \in \mathbb{R}^n, \quad (1-\varrho)\|x\|^2 \leq \frac{1}{m}\|Ax\|^2\right\} \tag{3.3.5}$$
represents injectivity of the measurement matrix A.
- $\mathcal{E}_{\text{smad}}$ is the event on which the function f is L-smooth relative to ψ in the sense of Definition 2.3.6, with $L = 3 + \varrho\max(\|\bar{x}\|^2/3, 1)$.
- $\mathcal{E}_{\text{scvx}}$ is the event on which f is σ-strongly convex on $B(\overline{\mathcal{X}}, \rho)$ relative to ψ in the sense of Definition 2.3.7, with $\sigma = (\lambda\min(\|\bar{x}\|^2, 1) - \varrho\max(\|\bar{x}\|^2/3, 1))$ and $\rho = \frac{1-\lambda}{\sqrt{3}}\|\bar{x}\|$.
- We end up by denoting
$$\mathcal{E}_{\text{conv}} = \mathcal{E}_{\text{strictsad}} \cap \mathcal{E}_{\text{conH}} \cap \mathcal{E}_{\text{inj}} \cap \mathcal{E}_{\text{smad}} \cap \mathcal{E}_{\text{scvx}}. \tag{3.3.6}$$

Our main result for Gaussian measurements is the following.

Theorem 3.3.3. *Fix $\lambda \in]0,1[$ and $\varrho \in]0, \lambda\min(\|\bar{x}\|^2, 1)/(2\max(\|\bar{x}\|^2/3, 1))[$. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1.*

(i) If the number of measurements m is large enough, i.e. $m \geq C(\varrho)n\log^3(n)$, then for almost all initializers x_0 of Algorithm 1 used with constant step-size $\gamma_k \equiv \gamma = \frac{1-\kappa}{3+\varrho\max(\|\bar{x}\|^2/3,1)}$, for any $\kappa \in]0,1[$, we have
$$\operatorname{dist}(x_k, \overline{\mathcal{X}}) \to 0,$$
and $\exists K \geq 0$, large enough such that $\forall k \geq K$,
$$\operatorname{dist}^2(x_k, \overline{\mathcal{X}}) \leq (1-\nu)^{k-K}\rho^2, \tag{3.3.7}$$
where
$$\nu = \frac{(1-\kappa)\left(\lambda\min(\|\bar{x}\|^2, 1) - \varrho\max(\|\bar{x}\|^2/3, 1)\right)}{3 + \varrho\max(\|\bar{x}\|^2/3, 1)}. \tag{3.3.8}$$
This holds with a probability at least $1 - 2e^{-\frac{m(\sqrt{1+\varrho}-1)^2}{8}} - 5e^{-\zeta n} - 4/n^2 - c/m$, where $C(\varrho), c$ and ζ are numerical positive constants.

(ii) Suppose moreover that ϱ obeys
$$\varrho \leq \eta_1^{-1}\left(\frac{1-\lambda}{\sqrt{3\left(6(1+(1-\lambda)^2/3)+1\right)}}\frac{1}{\max\left(\|\bar{x}\|, 1\right)}\right),$$
where η_1 is the function defined in (3.6.11). When $m \geq C(\varrho)n\log(n)$, if Algorithm 1 is initialized with the spectral method in Algorithm 2, then with probability at least $1 - 2e^{-\frac{m(\sqrt{1+\varrho}-1)^2}{8}} - 5e^{-\zeta n} - 4/n^2$ (ζ is a fixed numerical constant), (3.3.7) holds for all $k \geq K = 0$.

Before proving our result, the following remarks are in order.

Remark 3.3.4.

- In the regime of claim (i), when x_0 is chosen uniformly at random, Algorithm 1 provably converges to the true vector $\bar{x}$ up to a sign change. In this case any initialization strategy becomes superfluous, though the number of measurements required then is slightly (polylogarithmically) higher than with spectral initialization.

- In the regime of of claim (ii), one has to use a spectral initialization to find a good initial guess, from which mirror descent converges locally linearly to $\bar{x}$ up to global sign change.
- When the true vector norm is one, as assumed in many works, the convergence rate takes the simple form $\left(1 - \frac{(1-\kappa)(\lambda-\varrho)}{3+\varrho}\right) \leq \frac{2}{3} + O((1-\lambda) + \kappa + \varrho)$.
- The convergence rate $1-\nu$ as given in (3.3.7)-(3.3.8) can be slightly improved as we did in (3.2.6) (here we dropped the denominator in (3.2.6)). It is also important to point out that our convergence rate is independent from the dimension n of the signal. This is in contrast with the Wirtinger flow [53, 52], whose convergence rate is $(1 - \frac{\text{cst}}{n})$ and thus dimension-dependent. Such dependence was removed for the truncated Wirtinger flow with Gaussian measurements [63].

To close these remarks, we strongly believe that handling the geometry of the problem through the framework of mirror/Bregman gradient descent with a wisely chosen entropy/kernel ψ is a key for this better behaviour in our case.

Proof.

(i) Assume for this claim that $\mathcal{E}_{\text{conv}}$ holds true; we will show later that this is indeed the case *w.h.p* when the number of measurements is as large as prescribed. The proof then consists in combining Theorem 3.2.7 and the characterization of the structure of $\text{crit}(f)$.

- Global convergence of the iterates: under event $\mathcal{E}_{\text{inj}}$ (see (3.3.5)), the sequence $(x_k)_{k\in\mathbb{N}}$ generated by Algorithm 1 is bounded; see the discussion in Remark 3.2.8. Since $\mathcal{E}_{\text{smad}}$ holds, Theorem 3.2.7(i)-(ii) ensure that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to $x^\star \in \text{crit}(f)$ and the induced sequence $(f(x_k))_{k\in\mathbb{N}}$ converges to $f(x^\star)$.
- Since $\mathcal{E}_{\text{strictsad}}$ holds also, we have by Theorem 3.2.7-(iv) that for almost all initial points x_0, the sequence $(x_k)_{k\in\mathbb{N}}$ converges to an element of $\text{crit}(f)\backslash\text{strisad}(f) = \overline{\mathcal{X}}$. We assume *w.l.o.g* that $x_k \to \bar{x}$ whence $\|x_k - \bar{x}\| \to 0$, and $f(x_k) \to \min(f) = 0$. Therefore, for $\eta \leq \frac{\rho}{\sqrt{\max(\Theta(\rho),1)}}$, there exists $\exists K = K(\eta)$ such that,

$$\forall k \geq K, \|x_k - \bar{x}\| < \eta \text{ and } f(x_k) \in]0, \eta[, \tag{3.3.9}$$

  i.e. for $k \geq K$, x_k is in an f-attentive neighborhood of $\bar{x}$.
- Thanks to $\mathcal{E}_{\text{scvx}}$, f is σ-strongly convex on $B(\bar{x}, \rho)$ relative to ψ with σ and ρ as given in that event. It then follows from Theorem 3.2.7(iii) that, $\forall k > K$ and $\gamma_k \equiv \frac{(1-\kappa)}{3+\varrho\max(\|\bar{x}\|^2/3,1)}$, we have

$$\begin{aligned} D_\psi(\bar{x}, x_{k+1}) &\leq (1-\nu)\, D_\psi(\bar{x}, x_k) \\ &\leq (1-\nu)^{k-K}\, D_\psi(\bar{x}, x_K). \end{aligned}$$

  Moreover by (2.3.5) and 1-strong convexity of ψ, for all $k \geq K$

$$\begin{aligned} \text{dist}^2(x_k, \overline{\mathcal{X}}) \leq \|x_k - \bar{x}\|^2 \leq 2D_\psi(\bar{x}, x_k) &\leq (1-\nu)^{k-K}\, \Theta(\rho)\eta^2, \\ &\leq (1-\nu)^{k-K}\, \rho^2. \end{aligned}$$

To conclude this part of the proof we need to compute the probability that the event $\mathcal{E}_{\text{conv}}$ occurs. We have,

$$\begin{aligned} \mathcal{E}_{\text{conv}} &= \mathcal{E}_{\text{strictsad}} \cap \mathcal{E}_{\text{conH}} \cap \mathcal{E}_{\text{inj}} \cap \mathcal{E}_{\text{smad}} \cap \mathcal{E}_{\text{scvx}}, \\ &= \mathcal{E}_{\text{strictsad}} \cap \mathcal{E}_{\text{conH}} \cap \mathcal{E}_{\text{inj}}, \end{aligned}$$

since $\mathcal{E}_{\text{smad}} \subset \mathcal{E}_{\text{conH}}$ and $\mathcal{E}_{\text{scvx}} \subset \mathcal{E}_{\text{conH}}$ thanks to Lemma 3.6.5 and Lemma 3.6.6 respectively. Owing to Lemma 4.6.2, the event $\mathcal{E}_{\text{conH}}$ holds true with a probability at least $1 - 5e^{-\zeta n} - \frac{4}{n^2}$, where ζ is a fixed numerical constant, with the proviso that $m \geq C(\varrho)n\log(n)$.

On the other hand, Lemma 3.6.4 tells us that, when $m \geq \frac{16}{\varrho^2}n$, the event $\mathcal{E}_{\text{inj}}$ is true with a probability at least $1 - 2e^{-\frac{m(\sqrt{1+\varrho}-1)^2}{8}}$. The study of the critical points of the objective f, see [168, Theorem 2.2], shows that when $m \geq C(\varrho)n\log^3(n)$, the event $\mathcal{E}_{\text{strictsad}}$ holds true with a probability $1 - \frac{c}{m}$ (where c a fixed numerical constant). Using a union bound, $\mathcal{E}_{\text{conv}}$ occurs with the stated high probability provided that $m \geq C(\varrho)n\log(n)$ for a large enough numerical constant $C(\varrho)$.

(ii) The proof of this claim is similar to the last part of claim (i) except that now, we invoke Lemma 3.6.7-(iii) to see that with probability at least at least $1 - 2e^{-\frac{m(\sqrt{1+\varrho}-1)^2}{8}} - 5e^{-\zeta n} - \frac{4}{n^2}$, the initial guess x_0 obtained by spectral initialization belongs to $B\left(\overline{\mathcal{X}}, \frac{\rho}{\sqrt{\max(\Theta(\rho),1)}}\right)$. We can now follow the reasoning in the last item of the proof of statement (i) to conclude.

□

3.3.3 CDP measurements

Our main result for the CDP measurements model is the following.

Theorem 3.3.5. *Let $\varrho \in]0,1[$ and $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1.*

(i) If the number of patterns P satisfies $P \geq C(\varrho)\log(n)$, then with a probability at least $1 - 1/n^2$, for almost all initializers x_0 of Algorithm 1 used with constant step-size $\gamma_k \equiv \gamma = \frac{1-\kappa}{L}$, for any $\kappa \in]0,1[$ and L given by Lemma 3.2.3, $(x_k)_{k\in\mathbb{N}}$ converges to an element in $\text{crit}(f)\backslash\text{strisad}(f)$.

(ii) Let $\delta \in]0, \min(\|\bar{x}\|^2, 1)/2[$. There exists $\rho_\delta > 0$ such that if ϱ is small enough (i.e. it satisfies (3.6.24)) and $P \geq C(\varrho)n\log^3(n)$, and if Algorithm 1 is initialized with the spectral method in Algorithm 2, then with probability at least $1 - \frac{4P+1}{n^3} - \frac{1}{n^2}$

$$\text{dist}^2(x_k, \overline{\mathcal{X}}) \leq \prod_{i=0}^{k-1}(1-\nu_i)\rho_\delta^2, \quad \forall k \geq 0, \tag{3.3.10}$$

where

$$\nu_i = \frac{(1-\kappa)\left(\min(\|\bar{x}\|^2, 1) - 2\delta\right)}{(1+\delta)L_i}. \tag{3.3.11}$$

Let us first discuss this result and compare it to the one for Gaussian measurements.

Remark 3.3.6.

- As far as global recovery guarantees are concerned, Theorem 3.3.5-(i) does not ensure exact recovery of $\pm\bar{x}$. This is in contrast with the Gaussian model where this was established in Theorem 3.3.3-(i). As we pointed out earlier in the introduction section, one of the main difficulties is that several of our arguments in the Gaussian case rely on uniform bounds, for instance on the Hessian and gradient, that need to hold simultaneously for all vectors $x \in \mathbb{R}^n$ *w.h.p* . Unfortunately, the CDP model enjoys much much less randomness to exploit in the mathematical analysis making this very challenging. Nevertheless, numerical evidence in the next section suggests that global exact recovery (without spectral initialization) holds for the CDP model as well.
- Theorem 3.3.5(ii) ensures local linear convergence to the true vectors $\pm\bar{x}$ when our algorithm is initialized with the spectral method. The convergence rate is expressed in terms of the step-sizes $\gamma_i = \frac{1-\kappa}{L_i}$, where the L_i's are expected to be much smaller than L in Lemma 3.2.3. It is tempting to use $2(1+\delta)^2$, the local relative smoothness constant in (3.6.20), as an upper-bound estimate of the L_i's. But one has to keep in mind that this is valid only locally on $B(\pm\bar{x}, \rho_\delta)$, and thus one

cannot use it when iterating from x_k to x_{k+1}. In our numerical experiments, we nevertheless observe that the linear convergence rate in (3.3.11) is well estimated by $\left(1 - \frac{(1-\kappa)\left(\min(\|\bar{x}\|^2,1)-2\delta\right)}{2(1+\delta)^3}\right)$. When $\|\bar{x}\| \leq 1$, this rate reads $\left(1 - \frac{(1-\kappa)(1-2\delta)}{2(1+\delta)^3}\right) \leq \frac{1}{2} + O(\kappa + \delta)$.

Proof.

(i) Under the bound on P, we know from Lemma 3.6.10 that the measurement operator A is injective with probability at least $1 - 1/n^2$. On this event, the objective f is coercive, and thus the sequence $(x_k)_{k\in\mathbb{N}}$ generated by Algorithm 1 is bounded. Since f is L-smooth relative to ψ according to Lemma 3.2.3, Theorem 3.2.7(i)-(ii) ensure that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to $x^\star \in \mathrm{crit}(f)$ and the induced sequence $(f(x_k))_{k\in\mathbb{N}}$ converges to $f(x^\star)$. Then using Theorem 3.2.7-(iv) we get the statement.

(ii) By Lemma 3.6.12-(iii), we have that the spectral initialization guess x_0 belongs to $B\left(\overline{\mathcal{X}}, \frac{\rho_\delta}{\sqrt{\max(\Theta(\rho_\delta),1)}}\right)$ with probability larger than $1 - \frac{4P+1}{n^3} - \frac{1}{n^2}$. Moreover, we know from Lemma 3.6.11 that with probability at least $1 - \frac{4P+1}{2n^3}$, f is σ-strongly convex on $B(\overline{\mathcal{X}}, \rho_\delta)$ relative to ψ with $\sigma = \frac{\left(\min(\|\bar{x}\|^2,1)-2\delta\right)}{1+\delta}$. The rest of the proof follows the same reasoning as in the last item of the proof of statement Theorem 3.3.3-(i). We omit the details.

□

3.4 Numerical Experiments

In this section, we discuss some numerical experiments to illustrate the efficiency of our phase recovery algorithm. We use the standard normal Gaussian and we consider the CDP model with a random ternary variable d, *i.e.* taking values in $\{-1, 0, 1\}$ with probability $\{1/4, 1/2, 1/4\}$. In each instance, we measured the relative error between the reconstructed vector $\tilde{x}$ and the true signal one $\bar{x}$ as

$$\frac{\mathrm{dist}(\tilde{x}, \overline{\mathcal{X}})}{\|\bar{x}\|}. \tag{3.4.1}$$

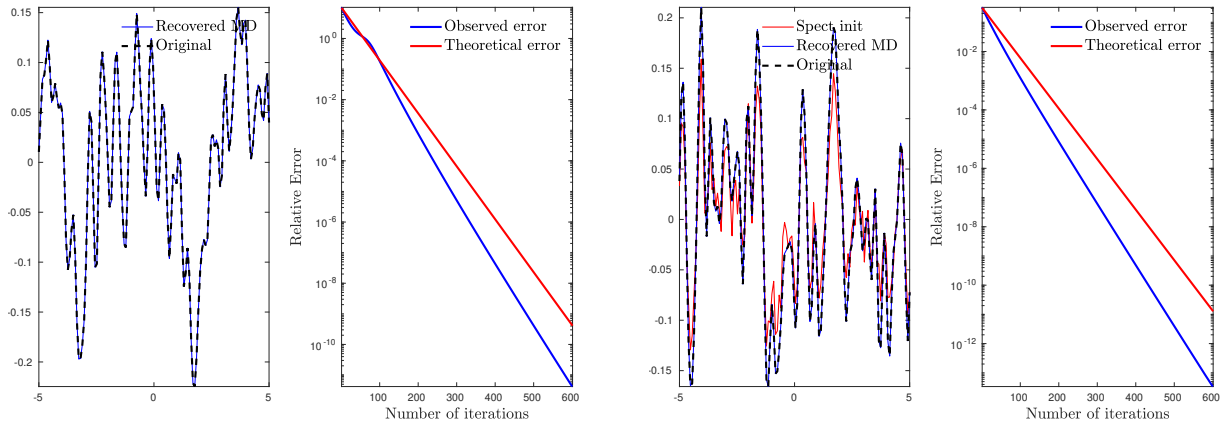
In the experiments, we set $\|\bar{x}\| = 1$ and $\tilde{x}$ was the output of Algorithm 1 at iteration K large enough.

3.4.1 Reconstruction of 1D signals

3.4.1.1 Gaussian measurements

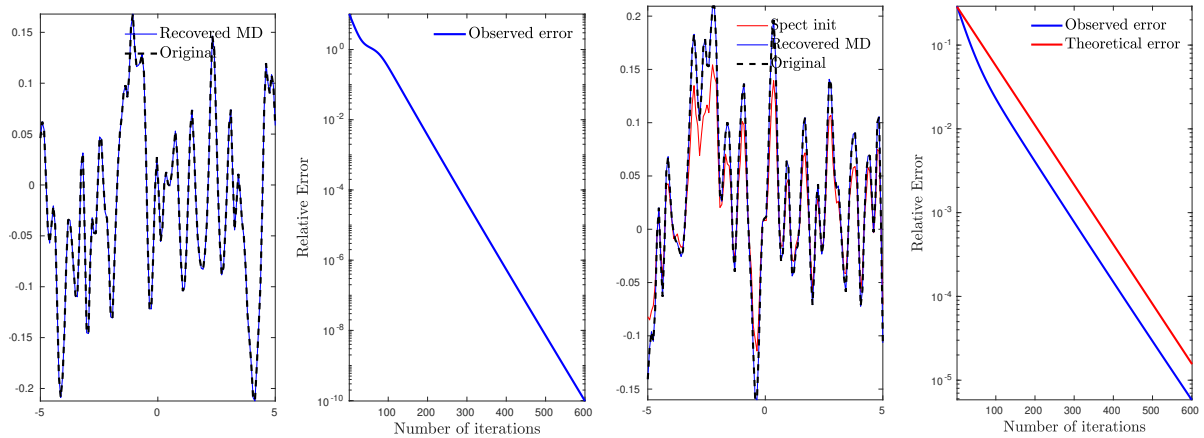
The goal is to recover a one-dimensional signal with $n = 128$ from Gaussian measurements. Figure 3.1(a) shows the reconstruction result from one random instance with $m = 2 \times 128 \times \log^3(128)$ without spectral initialization. Algorithm 1 was initialized with a vector drawn from the uniform distribution, and used with 600 iterations and a constant step-size $\gamma = \frac{0.99}{3}$. Given the oversampling rate, and as predicted by Theorem 3.3.3-(i), one can observe from Figure 3.1(a) that we have exact recovery, and after ~ 90 iterations, the iterates enter a linear convergence regime. The "Theoretical error" corresponds to the linear convergence rate predicted by (3.3.7)-(3.3.8), which is valid for k large enough.

Figure 3.1(b) displays the results for the case where $m = 2 \times 128 \times \log(128)$, and Algorithm 1 was applied with the same parameters as above except that the spectral initialization method was used to get the initial guess. As anticipated by Theorem 3.3.3-(ii), we again have exact recovery with a linear convergence behavior starting from the initial guess.



(a) Reconstruction with random initialization (b) Reconstruction with spectral initialization

Figure 3.1: Reconstruction of a 1D signal by mirror descent from Gaussian measurements.



(a) Reconstruction with random initialization (b) Reconstruction with spectral initialization

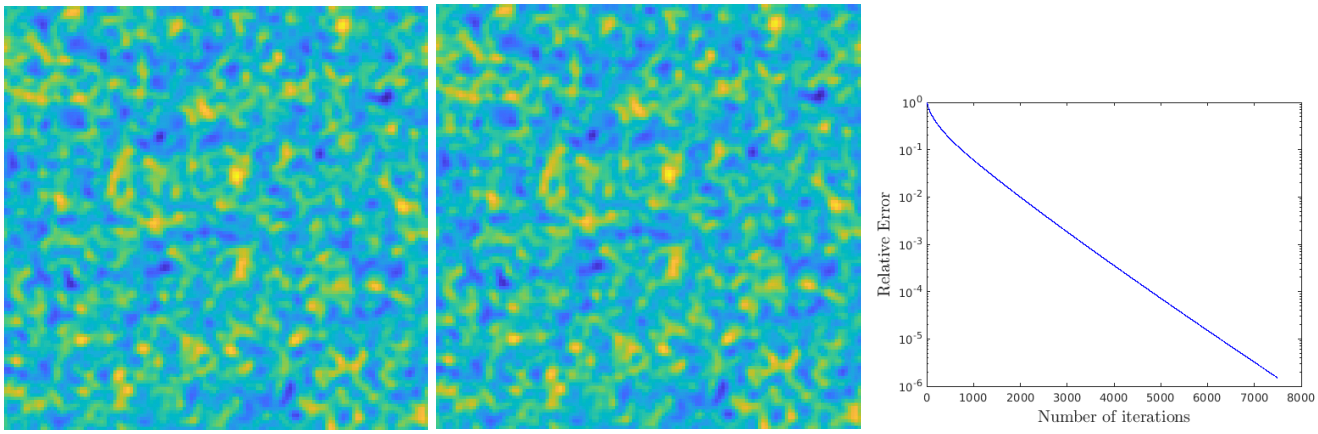
Figure 3.2: Reconstruction of a 1D signal by mirror descent from CDP measurements.

3.4.1.2 CDP measurements

We carried out the same experiment with the CDP measurements where we took $P = 7 \times \log^3(128)$ ternary random masks, and set $\gamma = \frac{0.99}{2}$ in mirror descent. The results are shown in Figure 3.2. The same conclusions drawn in the Gaussian case remain true for the CDP model. The results with spectral initialization depicted in Figure 3.2(b) are in agreement with those of Theorem 3.3.5-(ii). As for random uniform initialization, the results of Figure 3.2(a) provide numerical evidence that our algorithm enjoys global exact recovery properties, though this is so far not justified by our theoretical analysis.

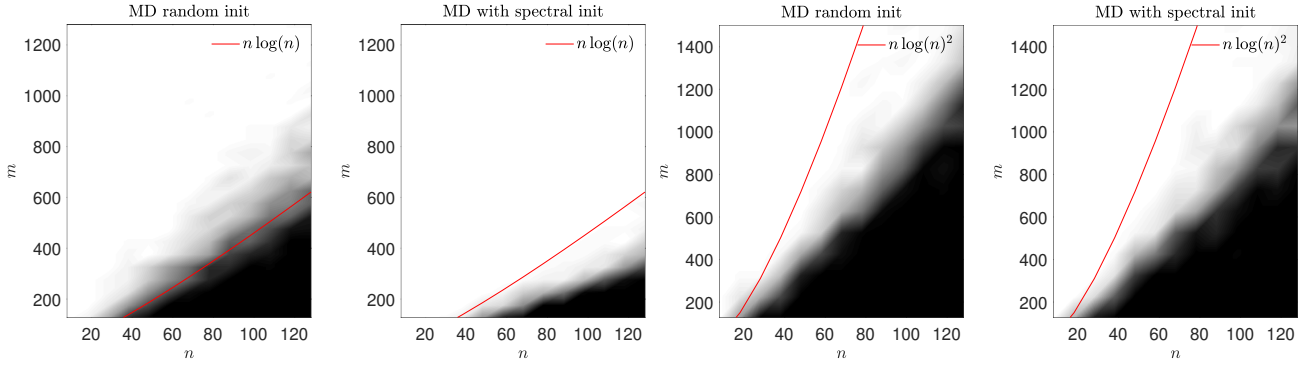
3.4.2 Recovery of the roughness of a 2D surface (light scattering)

In this experiment, we simulated a rough surface as a 256×256 Gaussian random field. The goal to recover this surface profile from the magnitude of the measurements according to the CDP model with $P = 100$ masks. The initial guess was drawn from the uniform distribution. The recovery results are displayed in Figure 3.3.



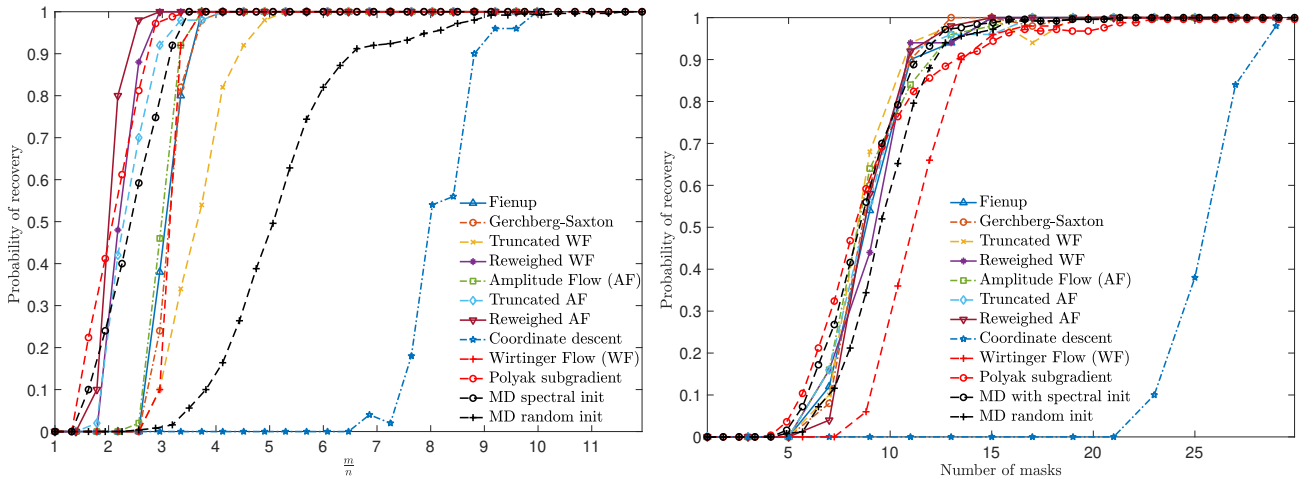
(a) Original surface (b) Recovered surface (c) Relative error

Figure 3.3: Roughness surface profile reconstruction by solving the phase retrieval problem from the CDP measurement model using mirror descent with uniform random initialization.



(a) Gaussian measurements (b) CDP measurements

Figure 3.4: Phase diagrams of mirror descent (MD) with spectral and uniform random initialization. (a) Gaussian measurements. (b) CDP measurements.



(a) Gaussian measurements (b) CDP measurements

Figure 3.5: Comparison of mirror descent to other methods in the literature. Each plot shows the empirical probability of success based on 100 random trials for two different measurement models (Gaussian and CDP) and a varied number of measurements.

3.4.3 Phase diagrams and comparison with other algorithms

Phase diagrams We first report the results of an experiment designed to estimate the phase retrieval probability for mirror descent, as we vary n and m. The results are depicted in Figure 3.4. For each pair (n, m), we generated 100 random instances and solved them with mirror descent (denoted MD for short hereafter), both with spectral initialization and with random uniform initialization. Each diagram shows the empirical probability (among the 100 random trials) that an algorithm successfully recovers the original vector up to a global sign change. We declared that a signal is recovered if the relative error (3.4.1) is less than 10^{-5}. The grayscale of each point in the diagrams reflects the empirical probability of success, from 0% (black) to 100% (white). The solid curve marks the prediction of the phase transition edge. One clearly sees a phase transition phenomenon which is in agreement with the predicted sample complexity bound shown as a solid line. For Gaussian measurements, MD with uniform random initialization has a transition to success occurring at a higher threshold compared to the version of MD with spectral initialization. This is in agreement with our theoretical findings. On the other hand, for CDP measurements, MD with uniform random initialization shows comparable performance to the version with spectral initialization especially as the oversampling (number of masks) increases, confirming numerically that spectral initialization does not seem to be mandatory for MD with CDP measurements.

Comparison with other algorithms We have also carried out a comprehensive comparative study of mirror descent (MD) to the methods included in the PhasePack library [60], which provides a common interface for testing phase retrieval methods on empirical datasets. We have used their implementations and included in the comparison MD and the Polyak subgradient method used in [69]. For fair comparison, and except MD with uniform initialization, we used spectral initialization for all algorithms. The results are displayed Figure 3.5 where each plot shows the empirical probability of success of each algorithm based on 100 random trials for two different measurement models (Gaussian and CDP) and a varied number of measurements. We fixed $n = 128$ in this experiment. References for all other algorithms as denoted in the legend in PhasePack can be found in [60].

For Gaussian measurements, MD with spectral initialization is in the group of best performing methods (Reweighted WF, Reweighted AF, Truncated AF, Polyak subgradient, MD) which exhibit comparable performance, though MD and Polyak subgradient are slightly better for low sampling rates (less than 2), and Reweighted AF appears better for $m/n \in [2, 3]$. This first group clearly outperforms the others especially when oversampling is less than 3. This is followed by a second group (AF, Fineup, Gerchberg-Saxton and WF), then Truncated WF, MD with random initialization, and finally the Coordinate Descent method. As far CDP measurements are concerned, most algorithms perform similarly and MD with spectral initilization appears to be among the best ones. MD with uniform random initialization has a recovery performance rather close to those ones, and better than the Wirtinger flow even if the latter uses spectral initialization.

3.5 Proofs for the Deterministic Case

Let us start this section by recalling our objective function *i.e.*

$$\forall x \in \mathbb{R}^n, \quad f(x) = \frac{1}{4m} \sum_{r=1}^{m} \left(|a_r^* x|^2 - y[r] \right)^2 = \frac{1}{4m} \sum_{r=1}^{m} \left(|a_r^* x|^2 - |a_r^* \bar{x}|^2 \right)^2, \tag{3.5.1}$$

The following expressions give the gradients and Hessians of f and ψ that will be used throughout. For all $\forall x \in \mathbb{R}^n$, we have

$$\nabla f(x) = \frac{1}{m}\sum_{r=1}^{m}\left(|a_r^* x|^2 - |a_r^*\bar{x}|^2\right) a_r a_r^* x, \qquad \nabla^2 f(x) = \frac{1}{m}\sum_{r=1}^{m}\left(3|a_r^* x|^2 - |a_r^*\bar{x}|^2\right) a_r a_r^*, \tag{3.5.2}$$

$$\nabla\psi(x) = \left(\|x\|^2 + 1\right) x, \qquad \nabla^2\psi(x) = \left(\|x\|^2 + 1\right)\mathrm{Id} + 2xx^\top. \tag{3.5.3}$$

3.5.1 Proof of Lemma 3.2.3

Proof. Our proof is different from that of [40, Lemma 5.1] and gives a better estimate of L. Since y has positive entries, we have for all $x, u \in \mathbb{R}^n$,

$$\begin{aligned}\left\langle u, \nabla^2 f(x) u\right\rangle &= \frac{1}{m}\sum_{r=1}^{m}\left(3|a_r^* x|^2 - y[r]\right)|a_r^* u|^2 \\ &\le \frac{1}{m}\sum_{r=1}^{m} 3|a_r^* x|^2 |a_r^* u|^2 \\ &\le \|x\|^2 \|u\|^2 \frac{1}{m}\sum_{r=1}^{m} 3\|a_r\|^4.\end{aligned}$$

On the other hand,

$$\begin{aligned}\left\langle u, \nabla^2\psi(x) u\right\rangle &= \left(\|x\|^2 + 1\right)\|u\|^2 + 2|\langle x, u\rangle|^2 \\ &\ge \|x\|^2\|u\|^2\end{aligned}$$

Thus for any $L \ge \frac{1}{m}\sum_{r=1}^{m} 3\|a_r\|^4$, we have for all $x \in \mathbb{R}^n$

$$\nabla^2 f(x) \preceq L\nabla^2\psi(x). \tag{3.5.4}$$

We conclude by invoking Lemma 2.3.8 with $g = f$ and $\phi = L\psi$, and Proposition 2.3.5-(ii). □

The following lemma states a key inequality that will be the starting point of our proof. It has appeared in different forms in the literature; see [40, Lemma 4.1 and Remark 4.1] or [169, Lemma 4.1]. We hereafter include a self-contained proof that accounts for backtracking.

Lemma 3.5.1. *Let $(x_k)_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 1. Then $\forall x \in \mathbb{R}^n$*

$$D_\psi(x, x_{k+1}) + \gamma_k\left(f(x_{k+1}) - f(x)\right) \le D_\psi(x, x_k) - \kappa D_\psi(x_{k+1}, x_k) - \gamma_k D_f(x, x_k). \tag{3.5.5}$$

Proof. From the update of x_{k+1}, we have $\nabla\psi(x_k) - \nabla\psi(x_{k+1}) = \gamma_k\nabla f(x_k)$, and multiplying both sides by $x_{k+1} - x$, we get

$$\langle\nabla\psi(x_k) - \nabla\psi(x_{k+1}), x_{k+1} - x\rangle = \gamma_k\langle\nabla f(x_k), x_{k+1} - x\rangle. \tag{3.5.6}$$

Using the three-point identity (2.3.4), we have

$$D_\psi(x, x_k) - D_\psi(x, x_{k+1}) - D_\psi(x_{k+1}, x_k) = \gamma_k\langle\nabla f(x_k), x_{k+1} - x\rangle, \tag{3.5.7}$$

By the backtracking test, we have that f verifies the L_k-relative smoothness inequality (2.3.6) *w.r.t* ψ at (x_{k+1}, x_k), with constant $L_k \le L$, that is

$$\begin{aligned}f(x_{k+1}) - f(x_k) &\le \langle\nabla f(x_k), x_{k+1} - x_k\rangle + L_k D_\psi(x_{k+1}, x_k) \\ &= \langle\nabla f(x_k), x_{k+1} - x\rangle + \langle\nabla f(x_k), x - x_k\rangle + L_k D_\psi(x_{k+1}, x_k),\end{aligned} \tag{3.5.8}$$

Plugging (3.5.7) into (3.5.8), we arrive at

$$\begin{aligned}&\gamma_k\left(f(x_{k+1}) - f(x_k)\right) \\ &\le D_\psi(x, x_k) - D_\psi(x, x_{k+1}) - D_\psi(x_{k+1}, x_k) + \gamma_k\langle\nabla f(x_k), x - x_k\rangle + \gamma_k L_k D_\psi(x_{k+1}, x_k) \\ &\le D_\psi(x, x_k) - D_\psi(x, x_{k+1}) - (1 - \gamma_k L_k) D_\psi(x_{k+1}, x_k) + \gamma_k\langle\nabla f(x_k), x - x_k\rangle \\ &\le D_\psi(x, x_k) - D_\psi(x, x_{k+1}) - \kappa D_\psi(x_{k+1}, x_k) + \gamma_k\langle\nabla f(x_k), x - x_k\rangle.\end{aligned}$$

Therefore

$$\begin{aligned}\gamma_k\left(f(x_{k+1})-f(x)\right) &\leq D_\psi\left(x,x_k\right)-D_\psi\left(x,x_{k+1}\right)-\kappa D_\psi\left(x_{k+1},x_k\right)\\ &\quad+\gamma_k\left(f(x_k)-f(x)+\langle\nabla f(x_k),x-x_k\rangle\right)\\ &= D_\psi\left(x,x_k\right)-D_\psi\left(x,x_{k+1}\right)-\kappa D_\psi\left(x_{k+1},x_k\right)-\gamma_k D_f(x,x_k).\end{aligned}$$

□

3.5.2 Proof of Theorem 3.2.7

Proof.

(i)-(ii) The objective function f in (3.2.1) is a real polynomial, hence obviously semi-algebraic. It then follows that f satisfies the Kurdyka-Łojasiewicz (KL) property [121, 122]. Combining this with Lemma 3.5.1, which ensures that the sequence $(x_k)_{k\in\mathbb{N}}$ is a gradient-like descent sequence, and 1-strong convexity of the entropy ψ, the proof of (i)-(ii) are similar to those of [40, Proposition 4.1,Theorem 4.1] with slight modifications to handle backtracking.

(iii)-(a) The proof of this claim follows the same steps as the proof of [8, Theorem 2.12] using again that f is a continuous function which satisfies the KL property, that $\min f = 0$ and that $(x_k)_{k\in\mathbb{N}}$ is a gradient-like descent sequence thanks to Lemma 3.5.1.

(iii)-(b) We verify by induction that $x_k \in B(x^\star,\rho), \forall k\in\mathbb{N}$. Observe first that $x_0 \in B(x^\star,r)\subset B(x^\star,\rho)$ since $r \leq \frac{\rho}{\max(\sqrt{\Theta(\rho)},1)} \leq \rho$. Suppose now that for $k\geq 0$, $x_i\in B(x^\star,\rho)$ for all $i\leq k$. From Lemma 3.5.1 applied at $x=x^\star$, and the optimality of $x^\star$, we have

$$\begin{aligned}D_\psi\left(x^\star,x_{k+1}\right) &\leq D_\psi\left(x^\star,x_{k+1}\right)-\left(1-\gamma_k L\right)D_\psi\left(x_{k+1},x_k\right)-\gamma_k D_f(x^\star,x_k)\\ &\leq D_\psi\left(x^\star,x_k\right)-\gamma_k D_f(x^\star,x_k)\\ &\leq \left(1-\gamma_k\sigma\right)D_\psi\left(x^\star,x_k\right)\\ &\leq \prod_{i=0}^{k}(1-\gamma_i\sigma)D_\psi(x^\star,x_0)\leq D_\psi\left(x^\star,x_0\right),\end{aligned}\tag{3.5.9}$$

where we used the positivity of D_ψ and the relative strong convexity on $B(x^\star,\rho)$. Now invoking Proposition 2.3.5-(iv), we have

$$\begin{aligned}\|x_{k+1}-x^\star\|^2 \leq 2D_\psi\left(x^\star,x_k\right) &\leq 2\prod_{i=0}^{k}(1-\gamma_i\sigma)D_\psi(x^\star,x_0)\\ &\leq \Theta(\rho)\left\|x_0-x^\star\right\|^2 \leq \frac{\Theta(\rho)}{\max(\Theta(\rho),1)}\rho^2\leq\rho^2,\end{aligned}$$

which entails that $x_i\in B(x^\star,\rho)$ for all $i\leq k+1$ as desired.

To show (3.2.6), we use again Lemma 3.5.1, relative strong convexity on $B(x^\star,\rho)$, and (3.5.9) to get

$$D_\psi\left(x^\star,x_{k+1}\right)+\gamma_k\sigma D_\psi\left(x_{k+1},x^\star\right)\leq D_\psi\left(x^\star,x_{k+1}\right)+\gamma_k\left(f(x_{k+1})-f^\star\right)\leq\left(1-\gamma_k\sigma\right)D_\psi\left(x^\star,x_k\right).\tag{3.5.10}$$

Now Proposition 2.3.5-(iv) and 1-strong convexity of ψ tell us that

$$D_\psi\left(x^\star,x_{k+1}\right)\leq\Theta(\rho)D_\psi\left(x_{k+1},x^\star\right).\tag{3.5.11}$$

Combining (3.5.10), (3.5.11), 1-strong convexity of ψ and that $2D_\psi\left(x^\star,x_0\right)\leq\rho^2$, we get the claim.

(iv) We need the following lemma which is an extension of [109, Proposition 10] to the more general $L-$smooth case.

Lemma 3.5.2. *Let F be defined as in (3.2.4) then,*

(a) $\forall x \in \mathbb{R}^n, \det \mathrm{D}F(x) \neq 0$,

(b) $\mathrm{strisad}(f) \subset U_F \overset{\text{def}}{=} \{x \in \mathbb{R}^n : \ F(x) = x, \max_i |\lambda_i(\mathrm{D}F(x))| > 1\}$.

Proof of Lemma 3.5.2. Recall that $F(x) = (\nabla\psi)^{-1}(\nabla\psi(x) - \gamma\nabla f(x))$. Denote $G(x) \overset{\text{def}}{=} \nabla\psi(x) - \gamma\nabla f(x)$ so that $F(x) = (\nabla\psi)^{-1} \circ G(x)$.

(a) Since ψ is C^2 function, and thus $\nabla\psi$ is C^1, and as ψ is strongly convex, the inverse function theorem ensures that $(\nabla\psi)^{-1}$ is a local diffeomorphism [1]. Therefore to have $\det \mathrm{D}F(x) \neq 0$, it suffices to show that G is a local diffeomorphism *i.e.* $\forall x \in \mathbb{R}^n, \mathrm{D}G(x)$ is an invertible linear transformation. We have $\mathrm{D}G(x) = \nabla^2\psi(x) - \gamma\nabla^2 f(x)$, and the $L-$relative smoothness property of f *w.r.t* ψ (see (3.5.4) in the proof of Lemma 3.2.3) implies that

$$\mathrm{D}G(x) = \nabla^2\psi(x) - \gamma\nabla^2 f(x) \succeq (1 - \gamma L)\nabla^2\psi(x) = \kappa\nabla^2\psi(x) \succeq \kappa \mathrm{Id} \succ 0.$$

where we used 1-strong convexity of ψ and that $\gamma L = 1 - \kappa \in]0, 1[$.

(b) For $x_\star \in \mathrm{strisad}(f)$, we have $F(x_\star) = x_\star$ since $\mathrm{strisad}(f) \subset \mathrm{crit}(f)$. It remains to show that $\det \mathrm{D}F(x_\star)$ has an eigenvalue of magnitude greater than one. We have,

$$\begin{aligned}\mathrm{D}F(x_\star) &\overset{\text{(Chain rule)}}{=} \nabla^2\psi^{-1}(G(x_\star))\mathrm{D}G(x_\star),\\ &= \nabla^2\psi^{-1}(x_\star)\left(\nabla^2\psi(x_\star) - \gamma\nabla^2 f(x_\star)\right),\\ &= \mathrm{Id} - \gamma\nabla^2\psi(x_\star)^{-1}\nabla^2 f(x_\star).\end{aligned}$$

Denote for short $H_\psi = \nabla^2\psi(x_\star)$. We then have

$$H_\psi^{1/2}\mathrm{D}F(x_\star)H_\psi^{-1/2} = \mathrm{Id} - \gamma H_\psi^{-1/2}\nabla^2 f(x_\star)H_\psi^{-1/2}.$$

$H_\psi^{1/2}\mathrm{D}F(x_\star)H_\psi^{-1/2}$ is symmetric. Let $v' = H_\psi^{1/2}v$ with v a unit-norm eigenvector associated to a strictly negative eigenvalue of $\nabla^2 f(x_\star)$. By the Courant-Fisher min-max theorem, we have

$$\begin{aligned}\lambda_{\min}(H_\psi^{-1/2}\nabla^2 f(x_\star)H_\psi^{-1/2}) &\leq \left\langle v', H_\psi^{-1/2}\nabla^2 f(x_\star)H_\psi^{-1/2}v' \right\rangle\\ &= \left\langle v, \nabla^2 f(x_\star)v \right\rangle < 0.\end{aligned}$$

In turn, $1 - \gamma\lambda_{\min}(H_\psi^{-1/2}\nabla^2 f(x_\star)H_\psi^{-1/2}) > 1$ is an eigenvalue of $H_\psi^{1/2}\mathrm{D}F(x_\star)H_\psi^{-1/2}$. Since, $H_\psi^{1/2}\mathrm{D}F(x_\star)H_\psi^{-1/2}$ is similar to $\mathrm{D}F(x_\star)$, we conclude.

□

To show (iv), we combine claim (ii), Lemma 3.5.2 and the centre stable manifold theorem (see [109, Corollary 1]) which allows to conclude that $\{x_0 \in \mathbb{R}^n : \lim_{k\to\infty} F^k(x_0) \in \mathrm{strisad}(f)\}$ has measure zero. □

3.6 Proofs for Random Measurements

3.6.1 Gaussian measurements

In this section, we assume that the sensing vectors $(a_r)_{r\in[\![m]\!]}$ follow the i.i.d standard Gaussian model.

[1] Recall that we have already argued that ψ is a Legendre function and thus $\nabla\psi$ is a bijection from $\mathbb{R}^n$ to $\mathbb{R}^n$ with inverse $(\nabla\psi)^{-1} = \nabla\psi^*$; see [155, Theorem 26.5]

3.6.1.1 Expectation and deviation of the Hessian

The next lemma gives the expression of the expectation of $\nabla^2 f(x)$.

Lemma 3.6.1. ***(Expectation of the Hessian)*** *Under the Gaussian model, we have*

$$\mathbb{E}\left(\nabla^2 f(x)\right) = 3\left(2xx^\top + \|x\|^2 \operatorname{Id}\right) - 2\bar{x}\bar{x}^\top - \|\bar{x}\|^2 \operatorname{Id}. \tag{3.6.1}$$

Proof. In view of (3.5.2), it is sufficient to compute

$$\mathbb{E}\left(\frac{1}{m}\sum_{r=1}^m |{a_r}^\top x|^2 a_r {a_r}^\top\right).$$

Computing this expectation is standard using independence and a simple moment calculation, which gives

$$\mathbb{E}\left(\frac{1}{m}\sum_{r=1}^m |{a_r}^\top x|^2 a_r {a_r}^\top\right) = 2xx^\top + \|x\|^2 \operatorname{Id}. \tag{3.6.2}$$

□

We now turn our attention to the concentration of the Hessian of f around its mean. We start with following key lemma.

Lemma 3.6.2. *Fix $\varrho \in]0,1[$. If the number of samples obeys $m \geq C(\varrho) n \log n$, for some sufficiently large $C(\varrho) > 0$, then*

$$\left\| \frac{1}{m}\sum_{r=1}^m |{a_r}^\top x|^2 a_r {a_r}^\top - \left(2xx^\top + \|x\|^2 \operatorname{Id}\right)\right\| \leq \frac{\varrho}{3}\|x\|^2.$$

holds simultaneously for all $x \in \mathbb{R}^n$ with a probability at least $1 - 5e^{-\zeta n} - \frac{4}{n^2}$, where ζ is a fixed numerical constant.

Proof. We follow a similar strategy to that of [53, Section A.4]. By a homogeneity argument and isotropy of the Gaussian distribution, it is sufficient to establish the claim for $x = e_1$, *i.e.* that

$$\left\| \frac{1}{m}\sum_{r=1}^m |a_r[1]|^2 a_r {a_r}^\top - \left(2e_1{e_1}^\top + \operatorname{Id}\right)\right\| \leq \frac{\varrho}{3}. \tag{3.6.3}$$

Since the matrix in (3.6.3) is symmetric, its spectral norm can be computed via the associated quadratic form, and (3.6.3) amounts to showing that

$$V(v) \overset{\text{def}}{=} \left| \frac{1}{m}\sum_{r=1}^m |a_r[1]|^2 |a_r^\top v|^2 - \left(1 + 2v[1]^2\right)\right| \leq \frac{\varrho}{3}$$

for all $v \in \mathbb{S}^{n-1}$. The rest of the proof shows this claim.

Let $\widetilde{a}_r = (a_r[2], \ldots, a_r[n])$ and $\widetilde{v} = (v[2], \ldots, v[n])$. We rewrite

$$|a_r^\top v|^2 = \left(a_r[1]v[1] + \widetilde{a}_r^\top \widetilde{v}\right)^2 = (a_r[1]v[1])^2 + \left(\widetilde{a}_r^\top \widetilde{v}\right)^2 + 2a_r[1]v[1]\widetilde{a}_r^\top \widetilde{v}.$$

We plug this decomposition into $V(v)$ to get

$$\begin{aligned} V(v) &= \left| \frac{1}{m}\sum_{r=1}^m a_r[1]^4 v[1]^2 + \frac{1}{m}\sum_{r=1}^m a_r[1]^2 (\widetilde{a}_r^\top \widetilde{v})^2 + 2\frac{1}{m}\sum_{r=1}^m |a_r[1]|^3 v[1]\widetilde{a}_r^\top \widetilde{v} - \left(\|\tilde{v}\|^2 + 3v[1]^2\right)\right|, \\ &\leq \left| \frac{1}{m}\sum_{r=1}^m a_r[1]^4 - 3\right| v[1]^2 + \left| \frac{1}{m}\sum_{r=1}^m a_r[1]^2 - 1\right| \|\tilde{v}\|^2 + 2\left| \frac{1}{m}\sum_{r=1}^m |a_r[1]|^3 v[1]\widetilde{a}_r^\top \widetilde{v}\right| \\ &+ \left| \frac{1}{m}\sum_{r=1}^m a_r[1]^2 \left(\widetilde{a}_r^\top \widetilde{v} - \|\tilde{v}\|^2\right)\right|. \end{aligned}$$

If $X \sim \mathcal{N}(0,1)$ we have $\mathbb{E}\left(X^{2p}\right) = \frac{(2p)!}{2^p p!}$ for $p \in \mathbb{N}$, and in particular $\mathbb{E}\left(X^2\right) = 1$ and $\mathbb{E}\left(X^4\right) = 3$. By the Tchebyshev's inequality Proposition 2.6.5 and a union bound argument, $\forall \varepsilon > 0$, and a constant $C(\varepsilon) \approx \max\left(26, \frac{96}{\varepsilon^2}\right)$ such that when $m \geq C(\varepsilon)n$ we have,

$$\frac{1}{m}\sum_{r=1}^{m}\left(a_r[1]^4 - 3\right) < \varepsilon, \quad \frac{1}{m}\sum_{r=1}^{m}\left(a_r[1]^2 - 1\right) < \varepsilon, \quad \frac{1}{m}\sum_{r=1}^{m} a_r[1]^6 \leq 20$$
$$\text{and} \quad \max_{1\leq r\leq m} |a_r[1]| \leq \sqrt{10\log m}.$$

Each of these event happens with probability at least $1 - \frac{1}{n^2}$, and thus their intersection occurs with a probability at least $1 - \frac{4}{n^2}$. On this intersection event, we have

$$V(v) \leq \varepsilon(v[1]^2 + \|\widetilde{v}\|^2) + 2\left|\frac{1}{m}\sum_{r=1}^{m} a_r[1]^3 v[1]\widetilde{a}_r^\top\widetilde{v}\right| + \left|\frac{1}{m}\sum_{r=1}^{m} a_r[1]^2\left(\widetilde{a}_r^\top\widetilde{v} - \|\widetilde{v}\|^2\right)\right|.$$

On the one hand, by a Hoeffding-type inequality Proposition 2.6.6, we have

$$\forall \varrho' > 0, \quad \left|\frac{1}{m}\sum_{r=1}^{m} a_r[1]^3 v[1]\widetilde{a}_r^\top\widetilde{v}\right| < \varrho' |v[1]| \, \|\widetilde{v}\|^2,$$

with a probability $1 - ee^{-\zeta' n} \geq 1 - 3e^{-\zeta' n}$, when $m \geq C(\varrho')\sqrt{n\sum_{r=1}^{m} a_r[1]^6}$ with $C(\varrho') \approx \frac{1}{\varrho'^2}$ and $\zeta' > 2$ an absolute constant.
On the other hand, by Bernstein-type inequality Proposition 2.6.7, we have

$$\forall \varrho' > 0, \quad \left|\frac{1}{m}\sum_{r=1}^{m} a_r[1]^2\left(\widetilde{a}_r^\top\widetilde{v} - \|\widetilde{v}\|^2\right)\right| \leq \varrho' \, \|\widetilde{v}\|^2,$$

with a probability $1 - 2e^{-\zeta' n}$, when $m \geq C(\varrho')\left(\sqrt{n\sum_{r=1}^{m} a_r[1]^4} + n\max_{1\leq r\leq m} a_r[1]^2\right)$ with $C(\varrho') \approx \frac{1}{\varrho'^2}$.
Overall, for any $v \in \mathbb{S}^{n-1}$, we have with probability at least $1 - 5e^{-\zeta' n}$

$$V(v) \leq \varepsilon + 3\varrho'.$$

At this stage, we use a covering argument ([176, Lemma 5.4]) with an $\frac{1}{2}-$net whose cardinality is smaller than 5^n. Therefore, choosing $\varepsilon = \varrho'$ and $\varrho = 12\varrho'$ we get the claim where $\zeta = \zeta' - \log(5) > 0$ since $\zeta' > 2$ in the Hoeffding and Bernstein inequalities used above. □

Lemma 3.6.3. ***(Concentration of the Hessian)*** *Fix $\varrho \in]0,1[$. If the number of samples obeys $m \geq C(\varrho)n\log n$, for some sufficiently large constant $C(\varrho) > 0$, then*

$$\left\|\nabla^2 f(x) - \mathbb{E}\left(\nabla^2 f(x)\right)\right\| \leq \varrho\left(\|x\|^2 + \frac{\|\bar{x}\|^2}{3}\right) \tag{3.6.4}$$

holds simultaneously for all $x \in \mathbb{R}^n$ with a probability at least $1 - 5e^{-\zeta n} - \frac{4}{n^2}$, where ζ is a fixed numerical constant.

Proof. Recall $\nabla^2 f(x)$ from (3.5.2). By the triangle inequality and Lemma 3.6.1, we have

$$\begin{aligned}\left\|\nabla^2 f(x) - \mathbb{E}\left(\nabla^2 f(x)\right)\right\| \leq & 3\left\|\frac{1}{m}\sum_{r=1}^{m} |a_r{}^\top x|^2 a_r a_r{}^\top - \left(2xx^\top + \|x\|^2 \operatorname{Id}\right)\right\| \\ & + \left\|\frac{1}{m}\sum_{r=1}^{m} |a_r{}^\top \bar{x}|^2 a_r a_r{}^\top - \left(2\bar{x}\bar{x}^\top + \|\bar{x}\|^2 \operatorname{Id}\right)\right\|.\end{aligned}$$

The claim is then a consequence of Lemma 3.6.2. □

3.6.1.2 Injectivity of the measurement operator

The next result shows that when the number of measurements is large enough, the measurement matrix A (whose rows are the $a_r{}^\top$'s) is injective *w.h.p* .

Lemma 3.6.4. *Fix $\varrho \in]0,1[$. Assume that $m \geq \frac{16}{\varrho^2} n$. Then*

$$(1-\varrho)\left\|x\right\|^2 \leq \frac{1}{m}\left\|Ax\right\|^2 \leq (1+\varrho)\left\|x\right\|^2, \quad \forall x \in \mathbb{R}^n. \tag{3.6.5}$$

This happens with a probability at least $1-2e^{-mt^2/2}$ with $\frac{\varrho}{4} = t^2 + t$.

Proof. This is a consequence of very standard deviation inequalities on the singular values of Gaussian random matrices; see [55, Lemma 3.1] for a similar statement. □

3.6.1.3 Relative smoothness

For the Gaussian phase retrieval, we have the following refined dimension-independent estimate of the relative smoothness modulus, which is much better that the bound of Proposition 3.2.3.

Lemma 3.6.5. *Fix $\varrho \in]0,1[$. If the event $\mathcal{E}_{\mathrm{conH}}$ defined by (3.3.4) holds true then,*

$$D_f(x,z) \leq \left(3 + \varrho \max(\left\|\bar{x}\right\|^2/3, 1)\right) D_\psi(x,z), \qquad \forall x, z \in \mathbb{R}^n. \tag{3.6.6}$$

Proof. Using (3.3.4), Lemma 3.6.1 and (3.5.3), we have

$$\begin{aligned}
\forall x \in \mathbb{R}^n, \quad \nabla^2 f(x) &\preceq \mathbb{E}\left(\nabla^2 f(x)\right) + \varrho\left(\left\|x\right\|^2 + \frac{\left\|\bar{x}\right\|^2}{3}\right)\mathrm{Id}, \\
&\preceq 3\left(2xx^\top + \left\|x\right\|^2 \mathrm{Id}\right) - 2\bar{x}\bar{x}^\top - \left\|\bar{x}\right\|^2 \mathrm{Id} \\
&\quad + \varrho \max(\left\|\bar{x}\right\|^2/3, 1)\left(\left\|x\right\|^2 + 1\right)\mathrm{Id}, \\
&\preceq 3\left(2xx^\top + (\left\|x\right\|^2 + 1)\mathrm{Id}\right) + \varrho \max(\left\|\bar{x}\right\|^2/3, 1)\nabla^2\psi(x), \\
&= 3\nabla^2\psi(x) + \varrho \max(\left\|\bar{x}\right\|^2/3, 1)\nabla^2\psi(x).
\end{aligned} \tag{3.6.7}$$

We conclude by applying Lemma 2.3.8. □

3.6.1.4 Local relative strong convexity

The next proposition establishes strong convexity of f relative to ψ on a sufficiently small ball around $\overline{\mathcal{X}}$. In view of strong 1-convexity of ψ, our result also implies strong convexity on the same ball as shown in [53, 168].

Lemma 3.6.6. *Fix $\lambda \in]0,1[$ and $\varrho \in]0, \lambda \min(\left\|\bar{x}\right\|^2, 1)/(2\max(\left\|\bar{x}\right\|^2/3, 1))$. If the event $\mathcal{E}_{\mathrm{conH}}$ defined by (3.3.4) holds true then for all $x, z \in B\left(\bar{x}, \frac{1-\lambda}{\sqrt{3}}\left\|\bar{x}\right\|\right)$ and $x, z \in B\left(-\bar{x}, \frac{1-\lambda}{\sqrt{3}}\left\|\bar{x}\right\|\right)$,*

$$D_f(x,z) \geq \left(\lambda \min(\left\|\bar{x}\right\|^2, 1) - \varrho \max(\left\|\bar{x}\right\|^2/3, 1)\right) D_\psi(x,z). \tag{3.6.8}$$

Observe that if $\left\|\bar{x}\right\| = 1$ the above result has a simpler statement. In particular, ϱ must lie in $]0, \lambda[$, and the local relative strong convexity modulus is $\lambda - \varrho$ on a ball of radius $\frac{1-\lambda}{\sqrt{3}}$ around $\overline{\mathcal{X}}$.

Proof. We embark from (3.3.4) and Lemma 3.6.1 to infer that $\forall x \in \mathbb{R}^n$

$$\nabla^2 f(x) \succeq -\varrho\left(\left\|x\right\|^2 + \frac{\left\|\bar{x}\right\|^2}{3}\right)\mathrm{Id} + 3\left(2xx^\top + \left\|x\right\|^2\mathrm{Id}\right) - \left(2\bar{x}\bar{x}^\top + \left\|\bar{x}\right\|^2\mathrm{Id}\right) \tag{3.6.9}$$

$$\succeq -\varrho \max(\left\|\bar{x}\right\|^2/3, 1)\nabla^2\psi(x) + 3\left(2xx^\top + \left\|x\right\|^2\mathrm{Id}\right) - \left(2\bar{x}\bar{x}^\top + \left\|\bar{x}\right\|^2\mathrm{Id}\right). \tag{3.6.10}$$

We then obtain, for any $v \in \mathbb{S}^{n-1}$

$$v^\top\nabla^2 f(x)v + \varrho \max(\left\|\bar{x}\right\|^2/3, 1)v^\top\nabla^2\psi(x)v \geq 3\left(2\left(v^\top x\right)^2 + \left\|x\right\|^2\right) - \left(2\left(v^\top\bar{x}\right)^2 + \left\|\bar{x}\right\|^2\right).$$

Let $\rho > 0$ small enough, to be made precise later. Thus for any $x = \pm\bar{x} + \rho v$ we get

$$\begin{aligned}
&v^\top \nabla^2 f(x) v + \varrho \max(\|\bar{x}\|^2/3, 1) v^\top \nabla^2 \psi(x) v \\
&\geq 6\left(v^\top \bar{x}\right)^2 + 6\rho^2 \pm 12\rho v^\top \bar{x} + 3\|\bar{x}\|^2 \pm 6\rho v^\top \bar{x} + 3\rho^2 - 2\left(v^\top \bar{x}\right)^2 - \|\bar{x}\|^2 \\
&= 4\left(v^\top \bar{x}\right)^2 + 9\rho^2 \pm 18\rho v^\top \bar{x} + 2\|\bar{x}\|^2 .
\end{aligned}$$

From (3.5.3), we also have

$$v^\top \nabla^2 \psi(x) v = \|x\|^2 + 1 + 2\left(v^\top x\right)^2 = 2\left(v^\top \bar{x}\right)^2 + 3\rho^2 + \pm 6\rho v^\top \bar{x} + \|\bar{x}\|^2 + 1.$$

Consider first the case where $\|\bar{x}\| \geq 1$. We then get

$$\begin{aligned}
&v^\top \left(\nabla^2 f(x) - \left(\lambda - \varrho \max(\|\bar{x}\|^2/3, 1)\right) \nabla^2 \psi(x)\right) v \\
&\geq 2(2-\lambda)\left(v^\top \bar{x}\right)^2 + 3(3-\lambda)\rho^2 \pm 6(3-\lambda)\rho v^\top \bar{x} + (2-\lambda)\|\bar{x}\|^2 - \lambda \\
&= 2(2-\lambda)\left(v^\top \bar{x}\right)^2 + 3(3-\lambda)\rho^2 \pm 6(3-\lambda)\rho v^\top \bar{x} + 2(1-\lambda)\|\bar{x}\|^2 + \lambda(\|\bar{x}\|^2 - 1) \\
&\geq 2(2-\lambda)\left(v^\top \bar{x}\right)^2 + 3(3-\lambda)\rho^2 \pm 6(3-\lambda)\rho v^\top \bar{x} + 2(1-\lambda)\|\bar{x}\|^2 .
\end{aligned}$$

We claim that

$$\inf_{v \in \mathbb{S}^{n-1}} 2(2-\lambda)\left(v^\top \bar{x}\right)^2 + 3(3-\lambda)\rho^2 \pm 6(3-\lambda)\rho v^\top \bar{x} + 2(1-\lambda)\|\bar{x}\|^2 \geq 0$$

for ρ small enough. Let $v^\top \bar{x} = \alpha \|\bar{x}\|$, where $\alpha \in [-1, 1]$ and $\rho = \beta \|\bar{x}\|$. Thus

$$\begin{aligned}
&2(2-\lambda)\left(v^\top \bar{x}\right)^2 + 3(3-\lambda)\rho^2 \pm 6(3-\lambda)\rho v^\top \bar{x} + 2(1-\lambda)\|\bar{x}\|^2 = \\
&\qquad \left(2(2-\lambda)\alpha^2 + 3(3-\lambda)\beta^2 \pm 6(3-\lambda)\alpha\beta + 2(1-\lambda)\right)\|\bar{x}\|^2 .
\end{aligned}$$

Minimizing the last term for α and substituting back, we have after simple algebra that

$$2(2-\lambda)\alpha^2 + 3(3-\lambda)\beta^2 \pm 6(3-\lambda)\alpha\beta + 2(1-\lambda) \geq 2(1-\lambda) - \phi(\lambda)\beta^2.$$

where we set the function $\phi : t \in]0, 1[\mapsto \frac{36(3-t)^2}{8(2-t)} - 3(3-t) \in \mathbb{R}_+$. It can be easily shown that $\sup_{]0,1[} \phi(t) = \phi(1) = 12$. In turn, we have

$$2(2-\lambda)\alpha^2 + 3(3-\lambda)\beta^2 \pm 6(3-\lambda)\alpha\beta + 2(1-\lambda) \geq 0$$

since we assumed that $\rho \leq \frac{1-\lambda}{\sqrt{3}}\|\bar{x}\| \leq \frac{2(1-\lambda)}{\sqrt{\phi(\lambda)}}\|\bar{x}\|$.

Let us now turn to the case where $\|\bar{x}\| \leq 1$. We then have

$$\begin{aligned}
&v^\top \left(\nabla^2 f(x) - \lambda\left(\|\bar{x}\|^2 - \varrho \max(\|\bar{x}\|^2/3, 1)\right) \nabla^2 \psi(x)\right) v \\
&\geq 2(2-\lambda\|\bar{x}\|^2)\left(v^\top \bar{x}\right)^2 + 3(3-\lambda\|\bar{x}\|^2)\rho^2 \pm 6(3-\lambda\|\bar{x}\|^2)\rho v^\top \bar{x} + 2\|\bar{x}\|^2 - \lambda\|\bar{x}\|^4 - \lambda\|\bar{x}\|^2 \\
&\geq 2(2-\lambda\|\bar{x}\|^2)\left(v^\top \bar{x}\right)^2 + 3(3-\lambda\|\bar{x}\|^2)\rho^2 \pm 6(3-\lambda\|\bar{x}\|^2)\rho v^\top \bar{x} + 2(1-\lambda)\|\bar{x}\|^2 + \lambda\left(\|\bar{x}\|^2 - \|\bar{x}\|^4\right) \\
&\geq 2(2-\lambda\|\bar{x}\|^2)\left(v^\top \bar{x}\right)^2 + 3(3-\lambda\|\bar{x}\|^2)\rho^2 \pm 6(3-\lambda\|\bar{x}\|^2)\rho v^\top \bar{x} + 2(1-\lambda)\|\bar{x}\|^2 \\
&= \left(2(2-\lambda\|\bar{x}\|^2)\alpha^2 + 3(3-\lambda\|\bar{x}\|^2)\beta^2 \pm 6(3-\lambda\|\bar{x}\|^2)\alpha\beta + 2(1-\lambda)\right)\|\bar{x}\|^2 .
\end{aligned}$$

Arguing as in the first case, we have

$$2(2-\lambda\|\bar{x}\|^2)\alpha^2 + 3(3-\lambda\|\bar{x}\|^2)\beta^2 \pm 6(3-\lambda\|\bar{x}\|^2)\alpha\beta + 2(1-\lambda) \geq 2(1-\lambda) - \phi(\lambda\|\bar{x}\|^2)\beta^2.$$

Thus, the right hand side is non-negative since

$$\beta \leq \frac{1-\lambda}{\sqrt{3}} \leq \frac{2(1-\lambda)}{\sqrt{\phi(\lambda\|\bar{x}\|^2)}},$$

where we used that $\|\bar{x}\|^2 \leq 1$ in the argument of ϕ.
Overall, we have shown that

$$v^\top \left(\nabla^2 f(x) - \left(\lambda \min(\|\bar{x}\|^2, 1)/2 - \varrho \max(\|\bar{x}\|^2/3, 1) \right) \nabla^2 \psi(x) \right) v \geq 0$$

for all $v \in \mathbb{S}^{n-1}$ and $\rho \leq \frac{1-\lambda}{\sqrt{3}} \|\bar{x}\|$. We complete the proof by invoking Lemma 2.3.8 and convexity of the ball. □

3.6.1.5 Spectral initialization

We now show that the initial guess x_0 generated by spectral initialization (Algorithm 2) belongs to a small f-attentive neighborhood of $\overline{\mathcal{X}}$.

Lemma 3.6.7. *Fix $\varrho \in]0,1[$. If the number of samples obeys $m \geq C(\varrho) n \log n$, for some sufficiently large constant $C(\varrho) > 0$, then with probability at least $1 - 2e^{-\frac{m(\sqrt{1+\varrho}-1)^2}{8}} - 5e^{-\zeta n} - \frac{4}{n^2}$, where ζ is a fixed numerical constant, x_0 satisfies:*

(i) $\operatorname{dist}(x_0, \overline{\mathcal{X}}) \leq \eta_1(\varrho) \|\bar{x}\|$, *where*

$$\begin{aligned} \eta_1 :]0,1[&\to]0,1[\\ \varrho &\mapsto \left(\sqrt{2 - 2\sqrt{1-\varrho}} + \varrho/2 \right), \end{aligned} \tag{3.6.11}$$

which is an increasing function.

(ii) $f(x_0) \leq \left(3 + \varrho \max(\|\bar{x}\|^2/3, 1) \right) \frac{\Theta(\eta_1(\varrho)\|\bar{x}\|)}{2} \eta_1(\varrho)^2 \|\bar{x}\|^2$.

(iii) Besides, for $\lambda \in]0,1[$, if

$$\varrho \leq \eta_1^{-1} \left(\frac{1-\lambda}{\sqrt{3\left(6(1+(1-\lambda)^2/3)+1\right)}} \frac{1}{\max\left(\|\bar{x}\|, 1\right)} \right), \tag{3.6.12}$$

then with the same probability as above $x_0 \in B\left(\overline{\mathcal{X}}, \frac{\rho}{\max\left(\sqrt{\Theta(\rho)}, 1\right)} \right)$ where $\rho = \frac{1-\lambda}{\sqrt{3}} \|\bar{x}\|$.

Proof.

(i) Denote the matrix

$$Y = \frac{1}{m} \sum_{r=1}^{m} y[r] a_r a_r^\top = \frac{1}{m} \sum_{r=1}^{m} |a_r^\top \bar{x}|^2 a_r a_r^\top.$$

By Lemma 3.6.2, we have *w.h.p*

$$\|Y - \mathbb{E}(Y)\| \leq \varrho \|\bar{x}\|^2.$$

Let $\widetilde{x}$ be the eigenvector associated with the largest eigenvalue $\widetilde{\lambda}$ of Y such that $\|\widetilde{x}\| = \|\bar{x}\|$ (obviously $\widetilde{\lambda}$ is nonnegative since Y is semidefinite positive). Then,

$$\begin{aligned} \varrho \|\bar{x}\|^2 \geq \|Y - \mathbb{E}(Y)\| &\geq \|\bar{x}\|^{-2} \left| \widetilde{x}^\top \left(Y - 2\bar{x}\bar{x}^\top - \|\bar{x}\|^2 \operatorname{Id} \right) \widetilde{x} \right| \\ &= \|\bar{x}\|^{-2} \left| \widetilde{\lambda} \|\bar{x}\|^2 - 2(\widetilde{x}^\top \bar{x})^2 - \|\bar{x}\|^4 \right|. \end{aligned}$$

Hence

$$2(\widetilde{x}^\top \bar{x})^2 \geq \widetilde{\lambda} \|\bar{x}\|^2 - (1+\varrho) \|\bar{x}\|^4.$$

Moreover, using Lemma 3.6.2 again entails that *w.h.p*

$$\widetilde{\lambda} \|\bar{x}\|^2 \geq \bar{x}^\top Y \bar{x} \geq \bar{x}^\top \left(2\bar{x}\bar{x}^\top + \|\bar{x}\|^2 \operatorname{Id} \right) \bar{x} - \varrho \|\bar{x}\|^4 = (3-\varrho) \|\bar{x}\|^4.$$

Combining the last two inequalities, we get

$$(\widetilde{x}^\top \bar{x})^2 \geq (1-\rho) \|\bar{x}\|^4.$$

It then follows that

$$\operatorname{dist}(\widetilde{x}, \overline{\mathcal{X}}) \leq \sqrt{2 - 2\sqrt{1-\rho}}\, \|\bar{x}\|\,.$$

By definition of x_0 in Algorithm 2, $x_0 = \sqrt{m^{-1}\sum_r y[r]}\, \frac{\widetilde{x}}{\|\widetilde{x}\|}$, and thus *w.h.p*

$$\|x_0 - \widetilde{x}\| = \left|\sqrt{\frac{m^{-1}\sum_r y[r]}{\|\bar{x}\|^2}} - 1\right| \|\bar{x}\| = \left|\sqrt{\frac{m^{-1}\|A\bar{x}\|^2}{\|\bar{x}\|^2}} - 1\right| \|\bar{x}\| \leq \varrho/2\, \|\bar{x}\|\,,$$

where we used Lemma 3.6.4. In turn,

$$\operatorname{dist}(x_0, \overline{\mathcal{X}}) \leq \operatorname{dist}(\widetilde{x}, \overline{\mathcal{X}}) + \|x_0 - \widetilde{x}\| \leq \left(\sqrt{2 - 2\sqrt{1-\varrho}} + \varrho/2\right) \|\bar{x}\|\,.$$

(ii) Under our sampling complexity bound, event $\mathcal{E}_{\text{conH}}$ defined by (3.3.4) holds true *w.h.p* . It then follows from Lemma 3.6.5 applied at $\bar{x}$ and x_0, that

$$D_f(x_0, \bar{x}) \leq \left(3 + \varrho \max(\|\bar{x}\|^2/3, 1)\right) D_\psi(x_0, \bar{x}). \tag{3.6.13}$$

Since $f(\bar{x}) = 0$ and $\nabla f(\bar{x}) = 0$, we obtain from Proposition 2.3.5-(iv) that

$$\begin{aligned} f(x_0) &\leq \left(3 + \varrho \max(\|\bar{x}\|^2/3, 1)\right) D_\psi(x_0, \bar{x}) \\ &\leq \left(3 + \varrho \max(\|\bar{x}\|^2/3, 1)\right) \frac{\Theta(\eta_1(\varrho)\, \|\bar{x}\|)}{2} \eta_1(\varrho)^2\, \|\bar{x}\|^2\,. \end{aligned} \tag{3.6.14}$$

(iii) In view of (i), it is sufficient to show that $\eta_1(\varrho)\, \|\bar{x}\| \leq \frac{\rho}{\max\left(\sqrt{\Theta(\rho)}, 1\right)}$. Since from Proposition 2.3.5-(iv) (see also Remark 3.2.8) we have

$$\Theta(\rho) \leq 6(\|\bar{x}\|^2 + \rho^2) + 1 \leq \left(6(1 + (1-\lambda)^2/3) + 1\right) \max\left(\|\bar{x}\|^2, 1\right),$$

and η_1 is an increasing function, we conclude.

□

3.6.2 CDP model measurements

In this section, we assume that the sensing vectors $(a_r)_{r \in [\![m]\!]}$ follow the CDP model introduced in Section 3.3.1.

3.6.2.1 Expectation and deviation of the Hessian

Lemma 3.6.8. ***(Expectation of the Hessian)*** *Under the CDP measurement model, the following holds*

$$\mathbb{E}\left(\nabla^2 f(x)\right) = 3\left(xx^\top + \|x\|^2 \operatorname{Id}\right) - \bar{x}\bar{x}^\top - \|\bar{x}\|^2 \operatorname{Id}. \tag{3.6.15}$$

Proof. From [52, Lemma 3.1], we have

$$\forall x \in \mathbb{R}^n, \quad \mathbb{E}\left(\frac{1}{nP} \sum_{j,p=1}^{n,P} |f_j^* D_p x|^2 D_p f_j f_j^* D_p\right) = xx^\top + \|x\|^2 \operatorname{Id}. \tag{3.6.16}$$

Combining this with (3.5.2) yields the claim. □

Unlike the Gaussian model, it turns out that it is very challenging to concentrate the Hessian of f around its mean simultaneously for all vectors $x \in \mathbb{R}^n$ with non-trivial sampling complexity bounds. The main reason is that the CDP model does not have enough randomness to be used in the mathematical analysis. However, one can still do that for a fixed vector x. The next lemma gives the Hessian deviation at $\pm\bar{x}$.

Lemma 3.6.9. ***(Concentration of the Hessian)*** *Fix $\delta \in]0,1[$. If the number of patterns obeys $P \geq C(\delta)\log^3(n)$, then with a probability at least $1 - \frac{4P+1}{2n^3}$*

$$\left\|\nabla^2 f(\bar{x}) - \mathbb{E}\left(\nabla^2 f(\bar{x})\right)\right\| \leq \delta \left\|\bar{x}\right\|^2. \tag{3.6.17}$$

Proof. Let f_j^* be the rows of the discrete Fourier transform, *i.e.* $f_j[\ell] = e^{i\frac{2\pi j\ell}{n}}$. With a slight adaptation to the real case of the argument in [53, Section A.4.1], we deduce that

$$\left\|\frac{1}{nP}\sum_{j,p}^{n,P} |f_j^* D_p \bar{x}|^2 D_p f_j f_j^* D_p - \left(\bar{x}\bar{x}^\top + \left\|\bar{x}\right\|^2 \mathrm{Id}\right)\right\| \leq \frac{\delta}{2}\left\|\bar{x}\right\|^2, \tag{3.6.18}$$

provided that $P \geq C(\delta)\log^3(n)$ with a probability at least $1 - \frac{4P+1}{2n^3}$. Combining this with Lemma 3.6.8, we conclude. □

3.6.2.2 Injectivity of the measurement operator

We now establish that for m large enough, the measurement matrix A is injective *w.h.p* . Recall that the rows of A are the a_r^*'s.

Lemma 3.6.10. *Fix $\varrho \in]0,1[$. Assume that $P \geq C(\varrho)\log(n)$. Then with a probability at least $1 - 1/n^2$*

$$(1-\varrho)\left\|x\right\|^2 \leq \frac{1}{m}\left\|Ax\right\|^2 \leq (1+\varrho)\left\|x\right\|^2, \quad \forall x \in \mathbb{R}^n. \tag{3.6.19}$$

Proof. This is a consequence of the fact that

$$\left\|\frac{1}{m}A^*A - \mathrm{Id}\right\| \leq \varrho$$

with the claimed probability. Indeed, as for [52, Lemma 3.3], the covariance matrix $\frac{1}{m}A^*A$ is diagonal with i.i.d diagonal entries whose expectation is $\mathbb{E}\left(d^2\right) = 1$, and the statement follows from Hoeffding's inequality and a union bound. □

3.6.2.3 Local relative smoothness and relative strong convexity

We now turn to proving local relative smoothness and relative strong convexity near the true vectors. Unlike the Gaussian case, we only have a local version of relative smoothness. The reason behind this, as discussed above, is that it seems very hard to have a uniform concentration bound for the Hessian of f around its mean for the CDP model. To circumvent this, we use a continuity argument.

Lemma 3.6.11. *Fix $\delta \in]0, \min(\left\|\bar{x}\right\|^2, 1)/2[$. Suppose that (3.6.17) holds. Then there exists $\rho_\delta > 0$ such that for all $x, z \in B\left(\bar{x}, \rho_\delta\right)$ and $x, z \in B\left(-\bar{x}, \rho_\delta\right)$*

$$\frac{\left(\min(\left\|\bar{x}\right\|^2, 1) - 2\delta\right)}{1+\delta} D_\psi(x,z) \leq D_f(x,z) \leq 2(1+\delta)^2 D_\psi(x,z). \tag{3.6.20}$$

Observe that while in the Gaussian case, the ball radius on which relative strong convexity holds is fixed and explicit, for the CDP model, we only know it exists and it depends on δ.

Proof. We prove the claim for $\bar{x}$ and the same holds obviously around $-\bar{x}$. Using (3.6.17) and (3.5.3) gives

$$\begin{aligned}
\nabla^2 f(\bar{x}) &\preceq \mathbb{E}\left(\nabla^2 f(\bar{x})\right) + \delta\left\|\bar{x}\right\|^2 \mathrm{Id} \\
&= 2\bar{x}\bar{x}^\top + (2+\delta)\left\|\bar{x}\right\|^2 \mathrm{Id} \\
&\preceq (2+\delta)\left(2\bar{x}\bar{x}^\top + (\left\|\bar{x}\right\|^2 + 1)\mathrm{Id}\right), \\
&= (2+\delta)\nabla^2\psi(\bar{x}).
\end{aligned}$$

Again, from (3.6.17) and (3.5.3), we get

$$\nabla^2 f(\bar{x}) \succeq \mathbb{E}\left(\nabla^2 f(\bar{x})\right) - \delta \left\|\bar{x}\right\|^2 \mathrm{Id} \succeq 2(\bar{x}\bar{x}^\top + \left\|\bar{x}\right\|^2 \mathrm{Id}) - \delta\nabla^2\psi(\bar{x})\mathrm{Id}.$$

If $\|\bar{x}\| \geq 1$, we arrive at

$$\nabla^2 f(\bar{x}) \succeq 2\bar{x}\bar{x}^\top + (\left\|\bar{x}\right\|^2 + 1)\mathrm{Id} - \delta\nabla^2\psi(\bar{x})\mathrm{Id} = (1-\delta)\nabla^2\psi(\bar{x}).$$

If $\|\bar{x}\| \leq 1$, we have

$$\begin{aligned}\nabla^2 f(\bar{x}) - \left(\left\|\bar{x}\right\|^2 - \delta\right)\nabla^2\psi(\bar{x}) &\succeq 2\bar{x}\bar{x}^\top + 2\left\|\bar{x}\right\|^2 \mathrm{Id} - 2\left\|\bar{x}\right\|^2 \bar{x}\bar{x}^\top - \left\|\bar{x}\right\|^4 \mathrm{Id} - \left\|\bar{x}\right\|^2 \mathrm{Id} \\ &= 2(1-\left\|\bar{x}\right\|^2)\bar{x}\bar{x}^\top + (\left\|\bar{x}\right\|^2 - \left\|\bar{x}\right\|^4)\mathrm{Id} \succeq 0.\end{aligned}$$

Therefore

$$\left(\min(\left\|\bar{x}\right\|^2, 1) - \delta\right)\nabla^2\psi(\bar{x}) \preceq \nabla^2 f(\bar{x}) \preceq (2+\delta)\nabla^2\psi(\bar{x}). \tag{3.6.21}$$

Combining (3.6.21) with continuity of $\nabla^2 f$ and 1-strong convexity of ψ, $\exists \rho_\delta > 0$ such that $\forall x \in B(\bar{x}, \rho_\delta)$ we have

$$\left(\min(\left\|\bar{x}\right\|^2, 1) - 2\delta\right)\nabla^2\psi(\bar{x}) \preceq \nabla^2 f(\bar{x}) - \delta\mathrm{Id} \preceq \nabla^2 f(x) \preceq \nabla^2 f(\bar{x}) + \delta\mathrm{Id} \preceq 2(1+\delta)\nabla^2\psi(\bar{x}). \tag{3.6.22}$$

Continuity of $\nabla^2\psi$ and 1-strong convexity of ψ also yield that $\forall x \in B(\bar{x}, \rho_\delta)$

$$\nabla^2\psi(\bar{x}) \preceq \nabla^2\psi(x) + \delta\mathrm{Id} \preceq (1+\delta)\nabla^2\psi(x) \quad \text{and} \quad \nabla^2\psi(x) \preceq \nabla^2\psi(\bar{x}) + \delta\mathrm{Id} \preceq (1+\delta)\nabla^2\psi(\bar{x}). \tag{3.6.23}$$

Combining (3.6.22) and (3.6.23), we obtain that $\forall x \in B(\bar{x}, \rho_\delta)$,

$$\frac{\left(\min(\left\|\bar{x}\right\|^2, 1) - 2\delta\right)}{1+\delta}\nabla^2\psi(x) \preceq \nabla^2 f(x) \preceq 2(1+\delta)^2\nabla^2\psi(x).$$

Invoking Lemma 2.3.8 and convexity of the ball, we get the statement. □

3.6.2.4 Spectral initialization

We now show the analogue of Lemma 3.6.7 for the CDP measurement model.

Lemma 3.6.12. *Fix $\varrho \in]0,1[$. If the number of patterns obeys $P \geq C(\varrho) n \log^3(n)$, for some sufficiently large constant $C(\varrho) > 0$, then with probability at least $1 - \frac{4P+1}{n^3} - \frac{1}{n^2}$, x_0 satisfies:*

(i) $\operatorname{dist}(x_0, \overline{\mathcal{X}}) \leq \eta_1(\varrho)\left\|\bar{x}\right\|$, *where* $\eta_1(\varrho) = \left(\sqrt{2 - 2\sqrt{1-\varrho}} + \varrho/2\right)$.

Let $\delta \in]0, \min(\left\|\bar{x}\right\|^2, 1)/2[$ and ρ_δ is the neighborhood radius in Lemma 3.6.11. Suppose that ϱ is sufficiently small, i.e.

$$\varrho \leq \min\left(\delta, \eta_1^{-1}\left(\frac{\rho_\delta / \left\|\bar{x}\right\|}{\sqrt{6(\left\|\bar{x}\right\|^2 + \rho_\delta^2) + 1}}\right)\right). \tag{3.6.24}$$

Then, with the same probability as above,

(ii) $f(x_0) \leq 2(1+\delta^2)\frac{\Theta(\eta_1(\varrho)\|\bar{x}\|)}{2}\eta_1(\varrho)^2\left\|\bar{x}\right\|^2$;

(iii) $x_0 \in B\left(\overline{\mathcal{X}}, \frac{\rho_\delta}{\max\left(\sqrt{\Theta(\rho_\delta)}, 1\right)}\right)$.

Proof. The proof of this claim is similar to that of Lemma 3.6.7 for the Gaussian case, where we now invoke Lemma 3.6.9 and Lemma 3.6.10 for statement (i). For the last two claims, we also use Lemma 3.6.11 and that ϱ is small enough as prescribed.

□

Chapter 4

Stable Phase Retrieval with Mirror Descent

In this chapter, we aim to reconstruct an n-dimensional real vector from m phaseless measurements corrupted by an additive noise. We extend the noiseless framework developed in Chapter 3, based on mirror descent (or Bregman gradient descent), to deal with noisy measurements and prove that the procedure is stable to (small enough) additive noise. In the deterministic case, we show that mirror descent converges to a critical point of the phase retrieval problem, and if the algorithm is well initialized and the noise is small enough, the critical point is near the true vector up to a global sign change. When the measurements are i.i.d Gaussian and the signal-to-noise ratio is large enough, we provide global convergence guarantees that ensure that with high probability, mirror descent converges to a global minimizer near the true vector (up to a global sign change), as soon as the number of measurements m is large enough. The sample complexity bound can be improved if a spectral method is used to provide a good initial guess. We complement our theoretical study with several numerical results showing that mirror descent is both a computationally and statistically efficient scheme to solve the phase retrieval problem. The contributions of this chapter can be summarized as follows:

Main contributions of this chapter

- For almost all initializers, mirror descent converges to a critical point near the true vectors (up to sign ambiguity) where the objective has no direction of negative curvature.
- For i.i.d Gaussian measurements, and in the regime where the signal-to-noise ratio is large enough, we provide a complete geometric characterization of the landscape of the nonconvex objective provided that $m \gtrsim n \log^3(n)$.
- A global convergence to a point in $\operatorname{Argmin}(f)$, which is near $\bar{x}$ (up to sign ambiguity), as soon as the number of samples is large enough. If $m \gtrsim n \log(n)$, using a spectral initialization method, we provide a local convergence to a vector in the neighborhood of the target vector (up to sign ambiguity).

Contents

4.1 Introduction

4.1.1 Problem statement

Our focus in this chapter is phase retrieval with possibly noisy measurements. In real applications, the intensity measurements are not perfectly acquired. For instance, let us consider light scattering for precision in optics [7] which is our motivating application, where the goal is to describe the roughness of a polished surface. The latter is illuminated with a laser source, and the diffusion is measured by moving a detector. Then the power spectral density of the surface topography can be directly measured. However, during the acquisition process, different types of noise can corrupt the measurements such as photon noise, thermal noise, Johnson noise, *etc.*. Knowing the statistical model underlying the noise and the way it contaminates the measurements can prove useful to achieve robust reconstruction. The noise model can then be incorporated as the negative log-likelihood in the minimization objective. There are several noise models used in phase retrieval. One of them is the signal-dependent Poisson noise model which models the photon count noise. Another noise model is the (complex-valued) noise arising from multiple scattering, which can be modelled by the (complex) circularly-symmetric Gaussian distribution, and used to describe Rayleigh fading channels encountered in communication systems. Yet another source of noise the thermal one or the incoherent background noise.

In this manuscript, and similarly to [55, 70, 63], we will work with a generic additive noise model,

without any particular statistical assumption, in which the noisy phase retrieval problem reads

$$\begin{cases} \text{Recover } \bar{x} \in \mathbb{R}^n \text{ from the measurements } y \in \mathbb{R}^m \\ y[r] = |a_r^* \bar{x}|^2 + \epsilon[r], \quad r \in [\![m]\!], \end{cases} \tag{NoisyPR}$$

where $[r]$ is the r-th entry of the corresponding vector, and $\epsilon \in \mathbb{R}^m$ is the noise vector. Throughout the chapter, A is the $m \times n$ matrix with a_r^*'s as its rows.

Since $\bar{x}$ is real-valued, the best one can hope for is to ensure that $\bar{x}$ is uniquely determined from its intensities up to a global sign. Phase retrieval is in fact an ill-posed inverse problem in general, and even for $\epsilon = 0$, checking whether a solution to (NoisyPR) exists or not is known to be NP complete [157]. The situation is even more complicated in presence of noise. Thus, one of the major challenges is to design efficient recovery algorithms and find conditions on m, $(a_r)_{r\in[\![m]\!]}$ and ϵ which guarantee stable recovery in presence of noise. This is the goal we pursue in this chapter.

In this chapter, we cast the noise-aware phase retrieval problem (NoisyPR) as the smooth but nonconvex minimization problem

$$\min_{x\in\mathbb{R}^n} \left\{ f(x) \overset{\text{def}}{=} \frac{1}{4m} \sum_{r=1}^{m} \left(y[r] - |(Ax)[r]|^2 \right)^2 \right\}. \tag{4.1.1}$$

In fact, this is the same problem as in (4.2.1) studied in the noiseless case in Chapter 3. There, we proposed a mirror descent algorithm based on a suitably chosen entropy. In particular, we analyzed the case where the measurements were either i.i.d standard Gaussian measurements or drawn from the Coded Diffraction Pattern (CDP) model. It is our aim in this chapter to extend these results to the noisy case and prove stability guarantees for mirror descent to minimize (4.2.1).

4.1.2 Contributions and relation to prior work

Stability of phase retrieval to (small enough) noise has been studied by several authors with various measurement ensembles and reconstruction procedures; see the detailed review in Section 1.2. In this chapter, we claim that mirror descent to solve (4.2.1) is stable against sufficiently small additive noise. This in turn provides recovery error bounds of the noisy phase retrieval problem (NoisyPR). In the deterministic case, we show that for almost all initializers, mirror descent converges to a critical point near the true vectors (up to sign ambiguity) where the objective has no direction of negative curvature. In the random case, we consider i.i.d Gaussian measurements, and in the regime where the signal-to-noise ratio is large enough (see Assumption 4.3.1), we provide a complete geometric characterization of the landscape of the nonconvex objective provided that $m \gtrsim n \log^3(n)$. In turn, this allows us to describe the set of the critical points of f as the union of the strict saddle points and global minimizers of f. From this, we provide a global convergence to a point in $\operatorname{Argmin}(f)$, which is near $\bar{x}$ (up to sign ambiguity), as soon as the number of samples is large enough. If $m \gtrsim n \log(n)$, using a spectral initialization method, we provide a local convergence to a vector in the neighborhood of the target vector (up to sign ambiguity). By "near" we mean a reconstruction error that eventually scales as $O\left(\frac{\|\epsilon\|}{\sqrt{m}\|\bar{x}\|}\right)$ which matches the minimax optimal bounded established in [63, Theorem 3]. Compared to the Wirtinger flow and variants, our algorithm, by adapting to the geometry, offers an easier and dimension-independent choice of the parameters (in fact one, the descent-step size), and has global convergence guarantees.

Our results can be easily extended to sub-Gaussian measurements with minor changes. The case where a_r's are drawn from the CDP model is, however, far more challenging. Indeed, this model enjoys less randomness compared to the (sub-)Gaussian case and many of our arguments that require the uniformization of some bounds that are difficult to extend to the CDP model. Nevertheless, numerical experiments suggest that stable recovery still holds for our mirror descent algorithm with CDP measurements.

4.1.3 Chapter organization

The rest of the chapter is organized as follows. In Section 4.5, we recall the mirror descent algorithm with backtracking and establish its global and local convergence guarantees in the deterministic case. In Section 4.6, we sample complexity bounds with Gaussian measurements for our deterministic guarantees to hold with high probability. Section 4.4 describes the numerical experiments. The proofs of technical results are deferred to Section 4.5 and Section 4.6, while Section 4.7 studies the landscape of the noise-aware objective with Gaussian measurements.

4.2 Deterministic Stable Recovery

4.2.1 Mirror descent with backtracking

Observe that the objective in (NoisyPR) can be decomposed as

$$f(x) = \frac{1}{4m}\sum_{r=1}^{m}\Big(|(Ax)[r]|^2 - |(A\bar{x})[r]|^2 - \epsilon[r]\Big)^2. \tag{4.2.1}$$

As argued in Section 3.2, the objective f is $C^2(\mathbb{R}^n)$ and nonconvex (in fact only weakly convex). Moreover, its gradient is not Lipschitz continuous. However, using the strongly convex entropy (see Proposition 3.2.2 for its properties),

$$\psi(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2, \tag{4.2.2}$$

f turns out to be smooth relative to ψ.

Lemma 4.2.1. *Let f and ψ defined in (4.1.1)-(4.2.2). f is $L-$smooth relative to ψ on $\mathbb{R}^n$ for any $L \geq \frac{1}{m}\sum_{r=1}^{m}\|a_r\|^2\left(3\|a_r\|^2 + \|\epsilon\|_\infty\right)$.*

See Section 4.5.1 for the proof. This estimate of the modulus of relative smoothness L is crude and depends also on noise. This estimate will be largely improved for Gaussian measurements. Observe that we recover Lemma 3.2.3 in the noiseless case.

This relative smoothness property is the key motivation behind considering the framework of mirror descent or Bregman gradient descent. The mirror descent scheme with backtracking, already stated in Algorithm 1, is recalled in Algorithm 3 for convenience.

Algorithm 3: Mirror Descent for Phase Retrieval

Parameters: $L_0 = L$ (see Lemma 4.2.1), $\kappa \in]0,1[$, $\xi \geq 1$;
Initialization: $x_0 \in \mathbb{R}^n$;
for $k = 0, 1, \ldots$ **do**
  repeat
    $L_k \leftarrow L_k/\xi$, $\gamma_k = \frac{1-\kappa}{L_k}$;
    $x_{k+1} = F(x_k) = \nabla\psi^*\left(\nabla\psi(x_k) - \gamma_k\nabla f(x_k)\right)$;
  until $D_f(x_{k+1}, x_k) > L_k D_\psi(x_{k+1}, x_k)$;
  $L_k \leftarrow \xi L_k$, $\gamma_k = \frac{1-\kappa}{L_k}$;
  $x_{k+1} = F(x_k)$.

In Algorithm 3, ψ^* is the Legendre-Fenchel conjugate of ψ. The pair (f, ψ) defined in (4.1.1)-(4.2.2) satisfies [40, Assumptions A, B, C, D] and thus the mapping F in Algorithm 3 is well-defined and single-valued on $\mathbb{R}^n$. Moreover, the mirror step $\nabla\psi^*(z)$ can be computed easily as $\nabla\psi^*(z) = t^*z$, where t^* is the unique real positive root of the third-order polynomial $t^3\|z\|^2 + t - 1 = 0$; see Proposition 3.2.4.

4.2.2 Deterministic recovery guarantees by mirror descent

Before the deterministic result, we start by recalling the notion of strict saddles.

Definition 4.2.2 (Strict saddle point). A point $x_\star \in \mathrm{crit}(f)$ is a strict saddle point of f if $\lambda_{\min}(\nabla^2 f(x_\star)) < 0$. The set of strict saddle points of f is denoted $\mathrm{strisad}(f)$.

We now claim that mirror descent is stable against additive noise, as demonstrated in the following theorem.

Theorem 4.2.3. *Consider the noisy phase retrieval problem cast as (4.1.1). Let $(x_k)_{k\in\mathbb{N}}$ be a bounded sequence generated by Algorithm 3. Then,*

(i) the sequence $(x_k)_{k\in\mathbb{N}}$ has a finite length, converges to a point in $\mathrm{crit}(f)$ and the values $(f(x_k))_{k\in\mathbb{N}}$ are nonincreasing.
Take $L_k = L, \forall k \geq 0$. Then,

(ii) for Lebesgue almost all initializers x_0, the sequence $(x_k)_{k\in\mathbb{N}}$ converges to a critical point which cannot be a strict saddle, i.e. $x_k \to \widetilde{x} \in \mathrm{crit}(f)\backslash\mathrm{strisad}(f)$.

(iii) Assume that $\mathrm{Argmin}(f) \neq \emptyset$. Let $\rho, \sigma > 0$ such that $\rho > \frac{\sqrt{2}\|\epsilon\|}{\sqrt{m\sigma}}$ and define the radius $r \leq \sqrt{\frac{\rho^2 - \frac{2\|\epsilon\|^2}{m\sigma}}{\max(\Theta(\rho),1)}}$. If the initial point $x_0 \in B(\overline{\mathcal{X}}, r)$ and f is σ-strongly convex relative to ψ on $B(\overline{\mathcal{X}}, \rho)$ then $x_k \in B(\overline{\mathcal{X}}, \rho), \forall k \in \mathbb{N}$, and

$$\mathrm{dist}^2(x_k, \overline{\mathcal{X}}) \leq (1 - \gamma\sigma)^{k-1} \rho^2 + 2\frac{\|\epsilon\|^2}{m\sigma}, \tag{4.2.3}$$

See Section 4.5.2 for the proof.

Some remarks are in order.

Remark 4.2.4.

- Clearly, claim (i) suggests that even in the presence of the noise, any bounded sequence of Algorithm 3 will converge to a critical point of f with decreasing values. Let us observe that the sequence generated by our algorithm is bounded if for instance f is coercive, in which case $\mathrm{Argmin}(f)$ is also a non-empty compact set. This happens to be true when A is injective, *i.e.* in the oversampling regime as we will show in the random case.
- Claim (ii) shows that when the initial guess x_0 is chosen according to a distribution that has a density *w.r.t* the Lebesgue measure with constant step-size, then the sequence generated by mirror descent converges to a critical point where f has no direction of negative curvature.
- Concerning our local results in claim-(iii), if mirror descent is well initialized *i.e.*, in a ball of sufficiently small radius $r < \rho$ around the true vectors $\overline{\mathcal{X}}$, and if f is strongly convex relative to ψ on the larger ball $B(\overline{\mathcal{X}}, \rho)$, then all the iterates $(x_k)_{k\in\mathbb{N}}$ will remain in $B(\overline{\mathcal{X}}, \rho)$. Moreover, the sequence $(x_k)_{k\in\mathbb{N}}$ will converge to a critical point $\widetilde{x}$ obeying

$$\mathrm{dist}(\widetilde{x}, \overline{\mathcal{X}}) \leq \frac{\sqrt{2}\,\|\epsilon\|}{\sqrt{m\sigma}}.$$

4.3 Stable Recovery from Gaussian Measurements

The deterministic stable recovery results of Theorem 4.2.3(ii)-(iii) require for instance a local relative strong convexity condition around $\pm\bar{x}$ and possibly a good enough initial guess. A natural question to ask is when these conditions hold true. In turns out that this is indeed the case in the oversampling

regime with i.i.d Gaussian measurements, and if the noise is small enough. This section is devoted to rigorously show these statements.

We consider that the sensing vectors $(a_r)_{r\in[\![m]\!]}$ are drawn i.i.d from a real zero-mean standard Gaussian distribution. We also work under the following assumption on the noise ϵ.

Assumption 4.3.1. *Denote* $\widetilde{\epsilon} = \frac{1}{m}\sum_{r=1}^{m}\epsilon[r]$. *Given* $\lambda \in]0,1[$, *we suppose that*

$$0 \leq \frac{\widetilde{\epsilon}}{\min(\|\bar{x}\|^2, 1)} < \lambda \quad \textit{and} \quad \|\epsilon\|_\infty \leq c_s \min(\|\bar{x}\|^2, 1),$$
$$\textit{for some constant } c_s \in \left]0, \frac{(1-\lambda)\|\bar{x}\|\sqrt{\lambda\min(\|\bar{x}\|^2,1) - \widetilde{\epsilon}}}{2\sqrt{6}\min(\|\bar{x}\|^2,1)}\right[. \tag{4.3.1}$$

To get better understanding of this assumption, we observe that it implies that

$$\frac{\|\epsilon\|}{\sqrt{m}\|\bar{x}\|^2} \leq \frac{\|\epsilon\|_\infty}{\|\bar{x}\|^2} < \frac{(1-\lambda)\sqrt{\lambda}}{2\sqrt{6}} < 1.$$

On the other hand, for the observation model (NoisyPR) with i.i.d real Gaussian sensing vectors, the signal-to-noise ratio (SNR) is captured by

$$\mathrm{SNR} \overset{\text{def}}{=} \frac{\sum_{r=1}^{m}|a_r^\top \bar{x}|^4}{\|\epsilon\|^2} \approx \frac{3m\|\bar{x}\|^4}{\|\epsilon\|^2}.$$

In other words, Assumption 4.3.1 amounts to imposing that the SNR is large enough, *i.e.*

$$\sqrt{\mathrm{SNR}} \gtrsim \frac{6\sqrt{2}}{(1-\lambda)\sqrt{\lambda}}.$$

Let us also observe that Assumption 4.3.1 imposes that the empirical mean $\widetilde{\epsilon}$ is non-negative. This is a practical assumption that is fulfilled in many applications and will be helpful to describe the landscape of the noise-aware objective for Gaussian measurements. However, it was not used to have the deterministic guarantees.

Some of our stable recovery guarantees will be local provided that Algorithm 3 is initialized with a good guess. For this, we will use a spectral initialization method; see for instance [53, 63, 141, 192, 182, 179]. The procedure consists of taking x_0 as the leading eigenvector of a specific matrix as described in Algorithm 4.

Algorithm 4: Spectral Initialization.

Input: $y[r], r = 1, \ldots, m$

Output: x_0

Set $\lambda^2 = n\frac{\sum_r y[r]}{\sum_r \|a_r\|^2} = n\left(\frac{\sum_r \langle a_r, \bar{x}\rangle^2}{\sum_r \|a_r\|^2} + \frac{\sum_r \epsilon_r}{\sum_r \|a_r\|^2}\right)$;

Take x_0 the top eigenvector of $Y = \frac{1}{m}\sum_{r=1}^{m} y[r]a_r a_r^\top$ normalized to $\|x_0\| = \lambda$.

To lighten notations and clarify our proof, we consider the following events on whose intersection our deterministic convergence result will hold with high probability. Let fix $\varrho \in]0,1[$ and $\lambda \in \left]\frac{1}{9\sqrt{2}}, 1\right[$.

- The event
$$\mathcal{E}_{\text{strictsad}} = \{\mathrm{crit}(f) = \mathrm{Argmin}(f) \cup \mathrm{strisad}(f)\} \tag{4.3.2}$$
means that the set of critical points of the function f reduces to the set global minimizers of f and the set of strict saddle points.

- The event
$$\mathcal{E}_{\text{conH}} = \left\{\forall x \in \mathbb{R}^n, \quad \left\|\nabla^2 f(x) - \mathbb{E}\left(\nabla^2 f(x)\right)\right\| \leq \varrho\left(\|x\|^2 + \|\bar{x}\|^2/3 + \|\epsilon\|_\infty\right)\right\} \tag{4.3.3}$$

captures the deviation of the Hessian of f around its expectation.

- The event
$$\mathcal{E}_{\text{inj}} = \left\{ \forall x \in \mathbb{R}^n, \quad (1-\varrho)\|x\|^2 \leq \frac{1}{m}\|Ax\|^2 \right\} \tag{4.3.4}$$
corresponds to the injectivity of the measurement operator A.
- Let us denote by $\mathcal{E}_{\text{smad}}$ the event on which the function f is $L-$smooth relative to ψ with $L = 3 + \widetilde{\epsilon} + \varrho \max(\|\bar{x}\|^2/3 + 1, 1)$.
- Let us define $\rho = \frac{(1-\lambda)\|\bar{x}\|}{\sqrt{3}} > 0$ and $\mathcal{E}_{\text{scvx}}$ is the event on which f is σ-strongly convex relative to ψ locally on $B(\overline{\mathcal{X}}, \rho)$, with $\sigma = \lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon} - \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)$.
- We end up denoting
$$\mathcal{E}_{\text{conv}} = \mathcal{E}_{\text{strictsad}} \cap \mathcal{E}_{\text{conH}} \cap \mathcal{E}_{\text{inj}} \cap \mathcal{E}_{\text{smad}} \cap \mathcal{E}_{\text{scvx}}. \tag{4.3.5}$$

Our main result for Gaussian measurements is the following.

Theorem 4.3.2. *Fix* $\lambda \in \left]\frac{1}{9\sqrt{2}}, 1\right[$ *and* $\varrho \in \left]0, \frac{\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon}}{2\max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)}\right[$. *Let* $(x_k)_{k\in\mathbb{N}}$ *be the sequence generated by Algorithm 3. Under Assumption 4.3.1, let us define for any* $\kappa \in]0,1[$

$$\nu = \frac{(1-\kappa)\left(\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon} - \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)\right)}{3 + \widetilde{\epsilon} + \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)} \in [0,1[$$

and

$$\varsigma = \frac{2\sqrt{2}\|\epsilon\|}{\sqrt{m\left(c_s \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon}\right)}}.$$

(i) If the number of measurements m *is large enough i.e.,* $m \geq C(\varrho) n \log^3(n)$, *then for almost all initializers* x_0 *of Algorithm 3 with the step-size* $\gamma \equiv \frac{1-\kappa}{3+\widetilde{\epsilon}+\varrho\max(\|\bar{x}\|^2/3+\|\epsilon\|_\infty,1)}$, *then we have*

$$x_k \to x^\star \in \operatorname{Argmin}(f) \cap B\left(\overline{\mathcal{X}}, \varsigma\right)$$

and $\exists K > 0$ *such that for all* $k \geq K$, *we have*

$$\operatorname{dist}^2(x_k, \overline{\mathcal{X}}) \leq \frac{\|\bar{x}\|^2}{3}(1-\nu)^{k-K} + \varsigma^2. \tag{4.3.6}$$

This holds with a probability at least $1 - 2e^{-\frac{m(\sqrt{1+\varrho}-1)^2}{8}} - e^{-\Omega(m)} - 5e^{-\zeta n} - 4/n^2 - c/m$, *where* c, ζ *are fixed numerical constants.*

(ii) Suppose that ϱ *obeys (3.6.12). If* m *is such that* $m \geq C(\varrho, \|\epsilon\|_\infty) n \log(n)$, *and Algorithm 3 is initialized with the spectral method in Algorithm 4, then (4.3.6) holds for all* $k \geq K = 0$ *with probability at least* $1 - 2e^{-\frac{m(\sqrt{1+\varrho}-1)^2}{8}} - e^{-\Omega(m)} - 5e^{-\zeta n} - 4/n^2$, *where* ζ *is a fixed numerical constant.*

The choice of parameters can be made easier to read when $\|\bar{x}\| = 1$ as assumed in many works.

Corollary 4.3.3. *Suppose that* $\|\bar{x}\| = 1$ *and the noise is small enough. Fix* $\lambda \in \left]\frac{1}{9\sqrt{2}}, 1\right[$ *and* $\varrho \in \left]0, \frac{\lambda - \widetilde{\epsilon}}{2}\right[$. *Let* $(x_k)_{k\in\mathbb{N}}$ *be the sequence generated by Algorithm 3 with the step-size* $\gamma \equiv \frac{1-\kappa}{3+\widetilde{\epsilon}+\varrho}$, *where* $\kappa \in]0,1[$. *Then the statements of Theorem 4.3.2 hold true with*

$$\nu = \frac{(1-\kappa)(\lambda - \widetilde{\epsilon})}{3 + \widetilde{\epsilon} + \varrho} \quad \text{and} \quad \varsigma = \frac{2\sqrt{2}\|\epsilon\|}{\sqrt{m(c_s - \widetilde{\epsilon})}}.$$

Remark 4.3.4.

- When the number of measurements is sufficiently large as in claim (i), the SNR is large enough and the initial point x_0 is chosen randomly from a measure that has a density *w.r.t* Lebesgue measure, then mirror descent converges, eventually linearly, to an element of $\mathrm{Argmin}(f)$ which is within a factor of the noise level from $\overline{\mathcal{X}}$. The local convergence rate is dimension-independent. To the best of our knowledge, this is the first kind of results for the noisy phase retrieval problem.
- When the number of measurements is in the less demanding regime of the second claim, then mirror descent with spectral initialization again converges to a noise region around $\overline{\mathcal{X}}$.
- We recover the rate of Theorem 3.3.3 in the noiseless case.
- In the normalized setting of Corollary 4.3.3, the convergence rate behaves as $(1-\nu) \leq \frac{2}{3} + O\left((1-\lambda) + \kappa + \widetilde{\epsilon} + \varrho\right)$.
- It is important to observe that in the noisy case, the true vectors $\overline{\mathcal{X}}$ are not even critical points of f. Nonetheless, Lemma 4.5.2 will show that $\pm\bar{x}$ are actually $\frac{\|\epsilon\|^2}{m}$-minimizers.

Proof.

(i) We prove this claim by combining Theorem 4.2.3 and the characterization of the structure of $\mathrm{crit}(f)$ that we provide in Theorem 4.7.3. For the moment, let us assume that the event $\mathcal{E}_{\mathrm{conv}}$ holds true.

- By construction, $\mathcal{E}_{\mathrm{conv}} \subset \mathcal{E}_{\mathrm{inj}}$ which means that the operator A is injective showing the coercivity of the objective f which implies that the sequence $(x_k)_{k\in\mathbb{N}}$ generated by Algorithm 3 is bounded.
- From the event $\mathcal{E}_{\mathrm{smad}}$, we deduce that the function f is L-smooth relative to ψ with $L = 3 + \widetilde{\epsilon} + \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)$. Since the initializer is chosen at random with a fixed stepsize Theorem 4.2.3-(i)(ii) guarantees that $(x_k)_{k\in\mathbb{N}}$ converges to $x^\star \in \mathrm{crit}(f)\backslash\mathrm{strisad}(f)$ and $(f(x_k))_{k\in\mathbb{N}}$ also converges to $f(x^\star)$.
- The event $\mathcal{E}_{\mathrm{scvx}}$ shows that the function f is σ-strongly convex relative to ψ on $B(\overline{\mathcal{X}}, \rho)$ with $\sigma = \lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon} - \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)$. Given that Assumption 4.3.1 holds, Corollary 4.6.6 implies that $\rho^2 - \frac{4\|\epsilon\|^2}{m\sigma} > 0$ where we recall that $\rho = \frac{(1-\lambda)\|\bar{x}\|}{\sqrt{3}}$. Let us denote $r^2 = \frac{\rho^2 - \frac{4\|\epsilon\|^2}{m\sigma}}{\max(\Theta(\rho),1)}$.
- Moreover, $\mathcal{E}_{\mathrm{strictsad}}$ holds true, and thus $\mathrm{crit}(f)\backslash\mathrm{strisad}(f) = \mathrm{Argmin}(f)$, which means that for almost all initilizers x_0,

$$x_k \to x^\star \in \mathrm{Argmin}(f) \ \text{ and } \ f(x_k) - \min f \to 0. \tag{4.3.7}$$

We now claim that $\exists K > 0$, large enough, such that $\forall k \geq K, \mathrm{dist}(x_K, \overline{\mathcal{X}}) \leq r$ which will allow to invoke Theorem 4.2.3(iii). By Lemma 4.6.3, we have for any $k \in \mathbb{N}$

$$\mathrm{dist}(x_k, \overline{\mathcal{X}}) \leq \|x_k - x^\star\| + 8\frac{\|\epsilon\|}{\sqrt{m}\,\|\bar{x}\|}$$

with probability at least $1 - e^{-\Omega(m)}$. Since $x_k \to x_\star$, there exists K large enough such that $\forall k \geq K$

$$\mathrm{dist}(x_k, \overline{\mathcal{X}}) \leq 9\frac{\|\epsilon\|}{\sqrt{m}\,\|\bar{x}\|}$$

with the same probability. To conclude, it is then sufficient to show that

$$9\frac{\|\epsilon\|}{\sqrt{m}\,\|\bar{x}\|} \leq r.$$

This is true for sufficiently high SNR, *i.e.* under our Assumption 4.3.1. Therefore we deduce

from Theorem 4.2.3-(iii) that the sequence $(x_k)_{k\geq K} \in B(\overline{\mathcal{X}}, \rho)$ and for $k \geq K$,

$$\begin{aligned}\operatorname{dist}^2(x_k, \overline{\mathcal{X}}) &\leq \left(1 - \frac{(1-\kappa)\sigma}{L}\right)^{k-1} \rho^2 + \frac{4\|\epsilon\|^2}{m\sigma}, \\ &\leq (1-\nu)^{k-K}\rho^2 + 4\frac{\|\epsilon\|^2}{m\sigma}, \\ &\leq (1-\nu)^{k-K}\rho^2 + \varsigma^2,\end{aligned}$$

where we have used (4.6.12) *i.e.*, $\sigma \geq \frac{c_s \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon}}{2}$ which implies that $\operatorname{dist}(x_k, \overline{\mathcal{X}}) \leq \varsigma$.
Let us now compute the probability that $\mathcal{E}_{\text{conv}}$ occurs. The events $\mathcal{E}_{\text{smad}}, \mathcal{E}_{\text{scvx}}$ are contained in $\mathcal{E}_{\text{conH}}$, see respectively Lemma 4.6.4 and Lemma 4.6.5. From Lemma 4.6.2, $\mathcal{E}_{\text{conH}}$ occurs with a probability $1 - 5e^{-\zeta n} - 4/n^2 - 2e^{-\frac{m(\sqrt{1+\varrho}-1)^2}{8}}$ as soon as $m \geq C(\varrho) n \log(n)$. Besides, close observation of the Hessian concentration (noisy part) highlights that it implies the injectivity of the measurements thus $\mathcal{E}_{\text{inj}}$ is also contained in $\mathcal{E}_{\text{conH}}$. Thanks to Theorem 4.7.3, the event $\mathcal{E}_{\text{stricsad}}$ holds with a probability $1 - c/m$ as soon as $m \geq C(\varrho) n \log^3(n)$. Finally, we conclude with a union bound that $\mathcal{E}_{\text{conv}}$ holds with a probability at least $1 - 2e^{-\frac{m(\sqrt{1+\varrho}-1)^2}{8}} - 5e^{-\zeta n} - 4/n^2 - c/m$ (ζ, c are a fixed numerical constant) for $m \geq C(\varrho) n \log^3(n)$, which complete the proof.

(ii) By Lemma 3.6.4, the operator A is injective entailing by the coercivity of f that the sequence $(x_k)_{k\in\mathbb{N}}$ is bounded. From Corollary 4.6.6, when the signal-to-noise coefficient c_s satisfies (4.3.1), r is well-defined. Lemma 4.6.8 shows that when ϱ obeys (3.6.12), the initial point x_0 given by Algorithm 4 is in the right $f-$attentive topology at the distance at most $r = \sqrt{\rho^2 - \frac{4\|\epsilon\|^2}{m\sigma}}$. Thanks to Lemma 4.6.5, ρ is the radius of the ball $B(\overline{\mathcal{X}}, \rho)$ where we have σ-strong convexity relative to ψ with $\sigma = \lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon} - \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)$. The last point to check before applying Theorem 4.2.3 comes from Lemma 4.6.4 which shows that f is $L-$smooth relative to ψ with $L = 3 + \widetilde{\epsilon} + \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)$. We deduce from Theorem 4.2.3 that

$$\begin{aligned}\operatorname{dist}^2(x_k, \overline{\mathcal{X}}) &\leq \left(1 - \frac{(1-\kappa)\sigma}{L}\right)^{k-1} \rho^2 + \frac{4\|\epsilon\|^2}{m\sigma}, \\ &\leq (1-\nu)^{k-K}\rho^2 + 4\frac{\|\epsilon\|^2}{m\sigma} \\ &\leq (1-\nu)^{k-K}\rho^2 + \varsigma^2.\end{aligned}$$

Let us observe that this statement is true only on the intersection of the above events, we call it $\mathcal{E}'_{\text{conv}}$. Let now compute the probability that $\mathcal{E}'_{\text{conv}}$ occurs. We conclude with an appropriate union bound similar to the previous one, taking into account the fact that the spectral initialization event is contained in $\mathcal{E}_{\text{conH}}$. Finally, the statement holds with a probability at least $1 - 2e^{-\frac{m(\sqrt{1+\varrho}-1)^2}{8}} - 5e^{-\zeta n} - 4/n^2$ (ζ is a fixed numerical constant) for $m \geq C(\varrho) n \log(n)$, which completes the proof.

□

4.4 Numerical Experiments

In this section, we discuss some experiments to illustrate and validate numerically the efficiency of our phase recovery algorithm. In each instance, we measured the relative error between the reconstructed vector $\tilde{x}$ and the true signal one $\bar{x}$ as

$$\frac{\operatorname{dist}(\tilde{x}, \overline{\mathcal{X}})}{\|\bar{x}\|}. \tag{4.4.1}$$

In the experiments, we set $\|\bar{x}\| = 1$ and $\tilde{x}$ was the output of Algorithm 3 at iteration K large enough.

4.4.1 Experiments with Gaussian sensing vectors

4.4.1.1 Reconstruction of 1D signals

The aim is to reconstruct a randomly generated one-dimensional signal of length $n = 128$ from m noisy observations where the sensing vectors were drawn i.i.d from the standard Gaussian ensemble. The noise vector ϵ is chosen uniform such that $\widetilde{\epsilon} = 10^{-5}$.

In Figure 4.1, the blue line shows the evolution of the objective (left) and the relative error (right) using Algorithm 3 with $m = 128 \times \log^2(128)$, where the initial point was drawn from the uniform distribution. The red line is the result of reconstruction from $m = 5 \times 128 \times \log(128)$ measurements where Algorithm 3 was intialized using the spectral method in Algorithm 4 (the top eigenvalue was computed with the power iteration method using 200 iterations). In both cases, we run Algorithm 3 with a constant step-size $\gamma = \frac{0.99}{3+10^{-5}}$ (see Theorem 4.3.2). As predicted by the latter, the curves in blue (*i.e.* with random initialization) have two regimes: a sublinear regime and then a local linear regime. Moreover, with spectral initialization (red curves), f and the relative error converge linearly at the same rate as in the local regime of the blue curves, hence confirming our theoretical findings. As anticipated also, both f and the relative error eventually stabilize at a plateau whose level is governed by the noise level.

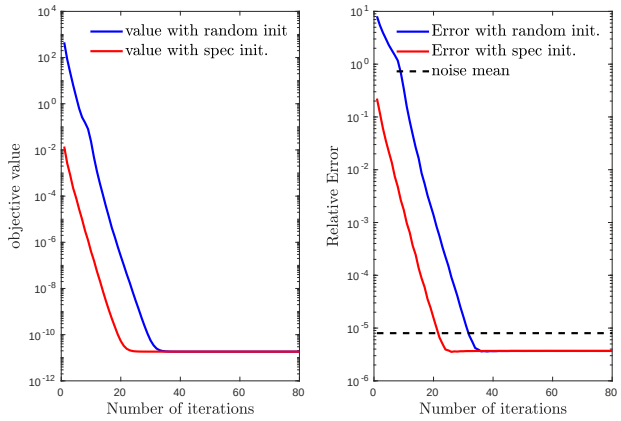


Figure 4.1: Reconstruction of signal from Gaussian measurements. The noise mean is $\widetilde{\epsilon}$.

4.4.1.2 Comparison with the Wirtinger flow

We compared our mirror descent algorithm (with and without spectral initialization) with the Wirtinger flow [53]. However, the Polyak subgradient method proposed in [69], that we included in our comparison in Chapter 3, is only applicable to the noiseless case as it needs the value of $\min f$ which is no longer known in presence of noise. We used the spectral method in Algorithm 4 for the Wirtinger flow, and we compared with mirror descent with and without spectral initialization.

For this, we report the results of an experiment designed to estimate the phase retrieval probability of success of each algorithm as we vary n and m. The results are depicted in Figure 4.2. For each pair (n, m), we generated 100 instances and solved them with each algorithm. Each diagram shows the empirical success probability (among the 100 instances) of the corresponding algorithm. An algorithm is declared as successful if the relative error (4.4.1) is less than $\frac{2\|\epsilon\|}{\sqrt{m}\sigma} \approx 10^{-5}$. The grayscale of each point in the diagrams reflects the observed probability of success, from 0% (black) to 100% (white). The solid curve marks the prediction of the phase transition edge. On the left panel of Figure 4.2, we also plot a profile of the phase diagram extracted at $n = 128$.

One observes a phase transition phenomenon that is in agreement with the predicted sample complexity bound shown as a solid line. Moreover, mirror descent performs better than the Wirtinger

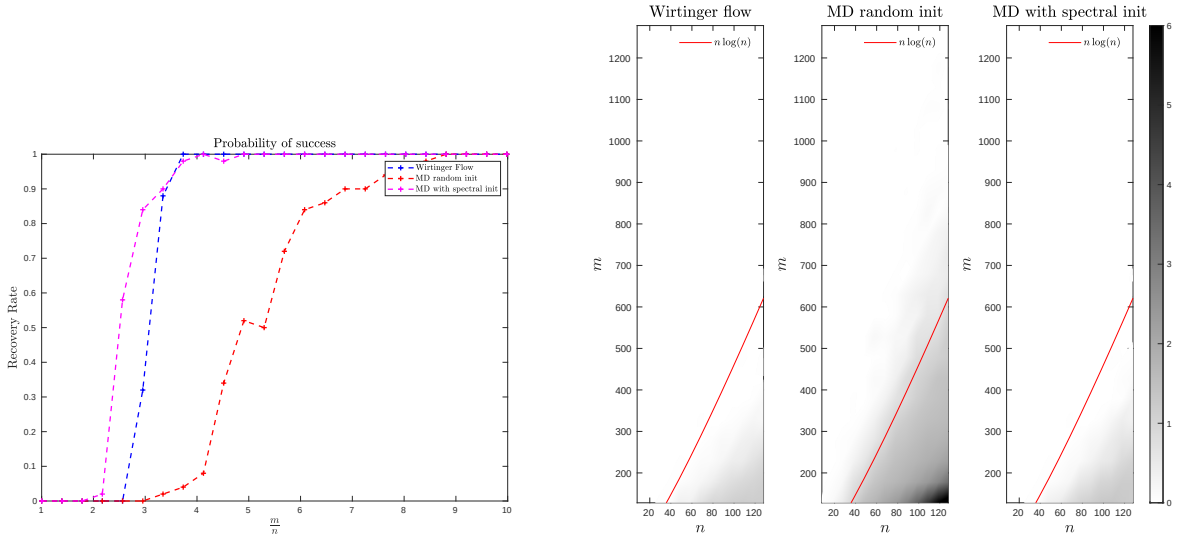


Figure 4.2: Phase diagrams for Gaussian measurements.

flow with both use spectral intialization. Mirror descent with uniform random initialization has a weaker recovery performance with a transition to success occurring at a higher threshold compared to the version of mirror descent with spectral initialization. This is in agreement with our theoretical findings as more measurements are needed in this case to ensure stable recovery.

4.4.2 Experiments with the CDP model

We now turn to the case of structured measurements from the CDP model. This model uses P coded diffraction patterns/masks followed by a Fourier transform. The observation model is given by

$$y = \left(\left| \sum_{\ell=0}^{n-1} \bar{x}_\ell d_p[\ell] e^{-i\frac{2\pi j\ell}{n}} \right|^2 + \epsilon_{j,p} \right)_{j,p}, \tag{4.4.2}$$

where $j \in \{0, \dots, n-1\}$ and $p \in \{0, \dots, P-1\}$, ϵ is the noise. The total number of measurements is thus $m = nP$ (*i.e.* the oversampling factor is P). $(d_p)_{p\in[P]}$ are i.i.d copies of a random variable d, and in our experiment, d takes values in $\{-1, 0, 1\}$ with probability $\{1/4, 1/2, 1/4\}$. Here we performed a similar experience to the Gaussian case described in Section 4.4.1.1, where we chose the number of masks $P = 7 \times \log^3(128)$ and a constant step-size $\gamma = \frac{0.99}{2+\widetilde{\epsilon}}$, with $\widetilde{\epsilon} = 10^{-5}$. Despite the lack of theoretical guarantees for the CDP model in the noisy case, that we conjecture are true, one can observe in Figure 4.3 that we have very similar results to those for Gaussian measurements.

4.4.3 Recovery of a 2D image

In this experiment, we work with the image of the beautiful Unicaen's[1] phoenix whose dimension is 396×396. Our goal is to recover the image from noisy CDP measurements with $P = 90$ masks. The noise is chosen such that $\widetilde{\epsilon} = 10^{-5}$. We used the spectral method to find the initial guess and run mirror descent for 1000 iterations. The results are displayed in Figure 4.4 showing that our algorithm converges to the desired image with a relative error of order 10^{-2}.

[1] Unicaen = University of Caen

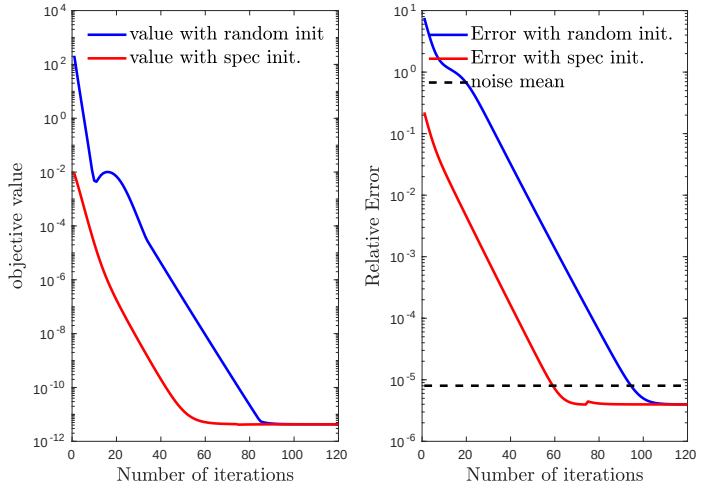
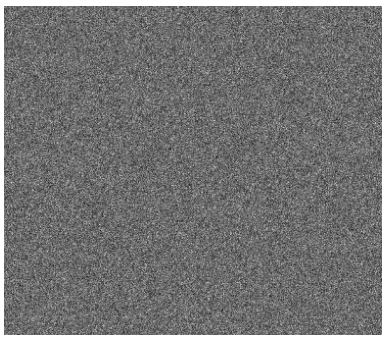


Figure 4.3: Reconstruction of signal from Noisy CDP. The noise mean is $\widetilde{\epsilon}$

(a) Original Unicaen's phoenix



(b) The CDP measurements averaged over the $P = 90$ masks

(c) Recovered Unicaen's phoenix

Figure 4.4: Reconstruction of an image from noisy CDP measurements.

4.5 Proofs for the Deterministic Case

Throughout the work, we use when it is convenient the following decomposition of the objective function f in (4.1.1).

$$\forall x \in \mathbb{R}^n, \quad f(x) = f_{\mathrm{NL}}(x) + f_{\mathrm{Ny}}(x), \tag{4.5.1}$$

where f_{NL} and f_{Ny} denote respectively the noiseless and the noisy part of f and we have

$$f_{\mathrm{NL}}(x) = \frac{1}{4m}\sum_{r=1}^{m}\left(|a_r{}^\top x|^2 - |a_r{}^\top \bar{x}|^2\right)^2, \qquad f_{\mathrm{Ny}}(x) = -\frac{1}{2m}\sum_{r=1}^{m}\epsilon[r]\left(|a_r{}^\top x|^2 - |a_r{}^\top \bar{x}|^2\right) + \frac{\|\epsilon\|^2}{4m}. \tag{4.5.2}$$

The following computations are straightforward:

$$\nabla\psi(x) = \left(\|x\|^2 + 1\right)x, \qquad \nabla^2\psi(x) = \left(\|x\|^2 + 1\right)\mathrm{Id} + 2xx^\top, \tag{4.5.3}$$

$$\nabla f(x) = \frac{1}{m}\sum_{r=1}^{m}\left(|a_r{}^\top x|^2 - |a_r{}^\top \bar{x}|^2\right)a_r a_r{}^\top x - \frac{1}{m}\sum_{r=1}^{m}\epsilon[r]a_r a_r{}^\top x \tag{4.5.4}$$

and

$$\nabla^2 f(x) = \frac{1}{m}\sum_{r=1}^{m}\left(3|a_r{}^\top x|^2 - |a_r{}^\top \bar{x}|^2\right)a_r a_r{}^\top - \frac{1}{m}\sum_{r=1}^{m}\epsilon[r]a_r a_r{}^\top. \tag{4.5.5}$$

4.5.1 Proof of Lemma 4.2.1

Proof. For all $x, u \in \mathbb{R}^n$, it easy to check that

$$\left\langle u, \nabla^2\psi(x)u\right\rangle \geq \left(\|x\|^2 + 1\right)\|u\|^2.$$

On the other hand, we have

$$\begin{aligned}
\left\langle u, \nabla^2 f(x)u\right\rangle &= \frac{1}{m}\sum_{r=1}^{m}\left(3|a_r{}^\top x|^2 - |a_r{}^\top \bar{x}|^2 - \epsilon[r]\right)|a_r{}^\top u|^2 \\
&\leq \frac{1}{m}\sum_{r=1}^{m}\left(3|a_r{}^\top x|^2 - \epsilon[r]\right)|a_r{}^\top u|^2 \\
&\leq \left(\frac{1}{m}\sum_{r=1}^{m}\left(3\|a_r\|^2 + \|\epsilon\|_\infty\right)\|a_r\|^2\right)\left(\|x\|^2 + 1\right)\|u\|^2.
\end{aligned}$$

Thus for any $L \geq \frac{1}{m}\sum_{r=1}^{m}\|a_r\|^2\left(3\|a_r\|^2 + \|\epsilon\|_\infty\right)$, we have for all $x \in \mathbb{R}^n$

$$\nabla^2 f(x) \preceq L\nabla^2\psi(x). \tag{4.5.6}$$

The claim then follows from Lemma 2.3.8 with $g = f$ and $\phi = L\psi$, and Proposition 2.3.5(ii). □

4.5.2 Proof of Theorem 4.2.3

Let us start with the following intermediate results.

Lemma 4.5.1. *Let the sequence $(x_k)_{k\in\mathbb{N}}$ be generated by Algorithm 3. Then for all $u \in \mathbb{R}^n$,*

$$D_\psi(u, x_{k+1}) + \gamma_k(f(x_{k+1}) - \min f) \leq D_\psi(u, x_k) - \kappa D_\psi(x_{k+1}, x_k) - \gamma_k D_f(u, x_k) + \gamma_k(f(u) - \min f). \tag{4.5.7}$$

Proof. From Lemma 3.5.1, we have

$$\forall u \in \mathbb{R}^n, \quad D_\psi(u, x_{k+1}) + \gamma_k\left(f(x_{k+1}) - f(u)\right) \leq D_\psi(u, x_k) - \kappa D_\psi(x_{k+1}, x_k) - \gamma_k D_f(u, x_k).$$

Subtracting $\min f$ from both sides yields the result. □

Lemma 4.5.2. *Assume that* $\operatorname{Argmin}(f) \neq \emptyset$. *We have*

$$0 \leq f(\pm\bar{x}) - \min f \leq \frac{\|\epsilon\|^2}{m}.$$

Proof. Let $x^\star \in \operatorname{Argmin}(f)$. We prove the claim for $\bar{x}$. By optimality, we have $f(x^\star) \leq f(\bar{x}) = \frac{\|\epsilon\|^2}{4m}$, which equivalently reads

$$\sum_{r=1}^m \left(|a_r{}^\top x^\star|^2 - |a_r{}^\top \bar{x}|^2\right)^2 \leq 2\sum_{r=1}^m \epsilon_r \left(|a_r{}^\top x^\star|^2 - |a_r{}^\top \bar{x}|^2\right).$$

Applying Young's inequality to the right-hand side then entails

$$(1-\delta)\sum_{r=1}^m \left(|a_r{}^\top x^\star|^2 - |a_r{}^\top \bar{x}|^2\right)^2 \leq \frac{\|\epsilon\|^2}{\delta} \quad \forall \delta \in]0,1[. \tag{4.5.8}$$

Consequently, using (4.5.8) and Young's inequality again, we get

$$\begin{aligned}
4mD_f(\pm\bar{x}, x^\star) &= 4m(f(\bar{x}) - f(x^\star)) \\
&= 2\sum_{r=1}^m \epsilon[r] \left(|a_r{}^\top x^\star|^2 - |a_r{}^\top \bar{x}|^2\right) - \sum_{r=1}^m \left(|a_r{}^\top x^\star|^2 - |a_r{}^\top \bar{x}|^2\right)^2 \\
&\leq 2\sum_{r=1}^m \epsilon[r] \left(|a_r{}^\top x^\star|^2 - |a_r{}^\top \bar{x}|^2\right) \\
&\leq \frac{\|\epsilon\|^2}{1-\delta} + (1-\delta)\sum_{r=1}^m \left(|a_r{}^\top x^\star|^2 - |a_r{}^\top \bar{x}|^2\right)^2 \leq \frac{\|\epsilon\|^2}{\delta(1-\delta)}.
\end{aligned}$$

The minimal value of the right hand side is $4\|\epsilon\|^2$ attained for $\delta = 1/2$. □

Proof.

(i)-(ii) Similar to the proofs of the corresponding claims in Theorem 3.2.7

(iii) We give the proof for $\bar{x}$ and obviously the same holds at $-\bar{x}$. We proceed by induction. We first have that $x_0 \in B(\bar{x}, r) \subset B(\bar{x}, \rho)$ since $r \leq \rho$. Suppose now that for $k \geq 0$, $(x_i)_{0\leq i\leq k} \subset B(\bar{x}, \rho)$. Applying Lemma 4.5.1 at $\bar{x}$ and using Lemma 4.5.2, we have

$$\begin{aligned}
D_\psi(\bar{x}, x_{k+1}) &\leq D_\psi(\bar{x}, x_k) - \kappa D_\psi(x_{k+1}, x_k) - \gamma D_f(\bar{x}, x_k) + \gamma\frac{\|\epsilon\|^2}{m} \\
&\leq D_\psi(\bar{x}, x_k) - \gamma D_f(\bar{x}, x_k) + \gamma\frac{\|\epsilon\|^2}{m} \\
&\leq (1-\gamma\sigma) D_\psi(\bar{x}, x_k) + \gamma\frac{\|\epsilon\|^2}{m},
\end{aligned} \tag{4.5.9}$$

where we also used positivity of D_ψ and local σ-relative strong convexity of f since $x_k \in B(\bar{x}, \rho)$. Iterating the last inequality, we get

$$\begin{aligned}
D_\psi(\bar{x}, x_{k+1}) &\leq (1-\gamma\sigma)^{k+1} D_\psi(\bar{x}, x_0) + \gamma\frac{\|\epsilon\|^2}{m}\sum_{i=0}^k (1-\gamma\sigma)^i \\
&\leq (1-\gamma\sigma)^{k+1} D_\psi(\bar{x}, x_0) + \frac{\|\epsilon\|^2}{m\sigma}\left(1 - (1-\gamma\sigma)^{k+1}\right) \\
&\leq D_\psi(\bar{x}, x_0) + \frac{\|\epsilon\|^2}{m\sigma}.
\end{aligned} \tag{4.5.10}$$

It then follows from Proposition 2.3.5(iv) that

$$\|x_{k+1} - \bar{x}\|^2 \leq \|x_0 - \bar{x}\|^2 \Theta(\rho) + \frac{2\|\epsilon\|^2}{m\sigma} \leq r^2\Theta(\rho) + \frac{2\|\epsilon\|^2}{m\sigma} \leq \rho^2.$$

This shows (4.2.3). □

4.6 Proofs for Gaussian Measurements

4.6.1 Expectation and deviation of the Hessian

Lemma 4.6.1. ***(Expectation of the Hessian)*** *If the sensing vectors $(a_r)_{r\in[m]}$ are sampled following the Gaussian model then we have for any $x \in \mathbb{R}^n$,*

$$\mathbb{E}\left(\nabla^2 f(x)\right) = 3\left(2xx^\top + \|x\|^2 \operatorname{Id}\right) - 2\bar{x}\bar{x}^\top - \|\bar{x}\|^2 \operatorname{Id} - \widetilde{\epsilon}\operatorname{Id}. \tag{4.6.1}$$

Proof. The proof combines (4.5.5), Lemma 3.6.1 for the expectation of the first (*i.e.* noiseless part), and the last term comes the fact that the sensing vectors have zero mean and unit covariance. □

Lemma 4.6.2. ***(Concentration of the Hessian)*** *Fix $\varrho \in]0,1[$, if the number of samples obeys $m \geq C(\varrho) n \log n$, for some sufficiently large constant $C(\varrho) > 0$ then*

$$\left\|\nabla^2 f(x) - \mathbb{E}\left(\nabla^2 f(x)\right)\right\| \leq \varrho\left(\|x\|^2 + \frac{\|\bar{x}\|^2}{3} + \|\epsilon\|_\infty\right) \tag{4.6.2}$$

holds simultaneously for all $x \in \mathbb{R}^n$ with a probability at least $1 - 5e^{-\zeta n} - \frac{4}{n^2} - 2e^{-\frac{m(\sqrt{1+\varrho}-1)^2}{8}}$, where ζ is a fixed numerical constant.

Proof. By the triangle inequality, we have

$$\begin{aligned}\left\|\nabla^2 f(x) - \mathbb{E}\left(\nabla^2 f(x)\right)\right\| \leq &\left\|\frac{1}{m}\sum_{r=1}^m \left(3|a_r{}^\top x|^2 a_r a_r{}^\top - |a_r{}^\top \bar{x}|^2 a_r a_r{}^\top\right) - \left(6xx^\top + 3\|x\|^2 \operatorname{Id} - 2\bar{x}\bar{x}^\top - \|\bar{x}\|^2 \operatorname{Id}\right)\right\| \\ &+ \left\|\frac{1}{m}\sum_{r=1}^m \epsilon[r] a_r a_r{}^\top - \widetilde{\epsilon}\operatorname{Id}\right\|.\end{aligned}$$

The concentration of the first term has been proved for the noiseless case (see [83, Lemma B.3]) with a probability $1 - 5e^{-\zeta n} - \frac{4}{n^2}$. For the noisy part, we have

$$\left\|\frac{1}{m}\sum_{r=1}^m \epsilon[r] a_r a_r{}^\top - \widetilde{\epsilon}\operatorname{Id}\right\| \leq \frac{\|\epsilon\|_\infty}{m}\left\|\sum_{r=1}^m \left(a_r a_r{}^\top - \operatorname{Id}\right)\right\| = \frac{\|\epsilon\|_\infty}{m}\left\|A^\top A - \operatorname{Id}\right\|, \tag{4.6.3}$$

where A is the $m \times n$matrix whose $r-$th row is the vector $a_r{}^\top$. From Lemma 3.6.4, we get that for any $\varrho \in]0,1[$,

$$\left\|\frac{1}{m}\sum_{r=1}^m \epsilon[r] a_r a_r{}^\top - \widetilde{\epsilon}\operatorname{Id}\right\| \leq \varrho\|\epsilon\|_\infty$$

with a probability at least $1 - 2e^{-mt^2/2}$, with $\frac{\varrho}{4} = t^2 + t$ with $m \geq \frac{16}{\varrho^2} n$. We conclude by applying a simple union bound. □

4.6.2 Optimal solution near the true vector

Lemma 4.6.3. *Assume that Assumption 4.3.1 holds and that $m \geq cn$ where c is a positive numerical constant. Then for any $x^\star \in \operatorname{Argmin}(f)$,*

$$\operatorname{dist}(x^\star, \overline{\mathcal{X}}) \leq 8\frac{\|\epsilon\|}{\sqrt{m}\,\|\bar{x}\|} \tag{4.6.4}$$

holds with probability $1 - e^{-\Omega(m)}$.

Proof. Let us use the following notation: $X^\star = x^\star x^{\star\top}, \overline{X} = \bar{x}\bar{x}^\top, \bar{\epsilon} \overset{\text{def}}{=} \frac{\|\epsilon\|}{\sqrt{m}}$ and $\epsilon_0 \overset{\text{def}}{=} 4\bar{\epsilon}$.

By optimality of $x^\star$, we have $f(x^\star) \leq f(\bar{x}) = \frac{\|\epsilon\|^2}{m}$, which also implies that

$$\sum_{r=1}^m \left(|a_r^* x^\star|^2 - |a_r^* \bar{x}|^2\right)^2 \leq 2\sum_{r=1}^m \epsilon_r\left(|a_r^* x^\star|^2 - |a_r^* \bar{x}|^2\right).$$

Applying Young's inequality to the right-hand side then entails

$$\sum_{r=1}^{m} \left(|a_r^* x^\star|^2 - |a_r^* \bar{x}|^2\right)^2 \leq 4 \|\epsilon\|^2 . \tag{4.6.5}$$

Fix $\zeta \in]0,1[$. Using [63, Lemma 1], there are positive numerical constants C and C' such that if $m \gtrsim n\zeta^{-2}\log(1/\zeta)$, then with probability at least $1 - C'e^{-C\zeta^2 m}$, we have

$$\frac{1}{m}\sum_{r=1}^{m} \left(|a_r^* x^\star|^2 - |a_r^* \bar{x}|^2\right)^2 \geq 0.81(1-\zeta)^2 \left\|X^\star - \overline{X}\right\|_{\mathrm{F}}^2.$$

Thus, in view of (4.6.5), we the same probability, we have

$$\left\|X^\star - \overline{X}\right\|_{\mathrm{F}} \leq \frac{20}{9(1-\zeta)}\bar{\epsilon}. \tag{4.6.6}$$

Therefore taking $\zeta = 0.4$ in (4.6.6) one has

$$\left\|X^\star - \overline{X}\right\|_{\mathrm{F}} \leq \epsilon_0. \tag{4.6.7}$$

The rest of the proof is inspired by that of [55, Theorem 1.2]. Since $\|x^\star\|^2$ and $\|\bar{x}\|^2$ are the largest eigenvalues of the rank-one symmetric matrices $X^\star$ and $\overline{X}$, we have from Weyl's perturbation inequality of the eigenvalues that

$$\left|\|x^\star\|^2 - \|\bar{x}\|^2\right| \leq \left\|X^\star - \overline{X}\right\|_{\mathrm{F}} \leq \epsilon_0.$$

Let us assume that $\|\bar{x}\|^2 = 1$ and the general case is obtained via a simple rescaling argument. Under Assumption 4.3.1, $\bar{\epsilon}$ is small enough so that $\epsilon_0 < 1$. We then get that $\|x^\star\|^2 \in [1-\epsilon_0, 1+\epsilon_0]$. The sin-$\theta$-Theorem [67] implies that

$$|\sin\theta| \leq \frac{\left\|X^\star - \overline{X}\right\|_{\mathrm{F}}}{\|x^\star\|^2} \leq \frac{\epsilon_0}{1-\epsilon_0},$$

where $0 \leq \theta \leq \frac{\pi}{2}$ is the angle between $x^\star$ and $\bar{x}$ which are the eigenvectors of $X^\star$ and $\overline{X}$ associated to the eigenvalues $\|x^\star\|^2$ and 1, respectively. We can then write

$$x^\star = \|x^\star\|\left(\cos\theta\bar{x} + \sin\theta\bar{x}^\perp\right),$$

where $\bar{x}^\perp$ is a unit vector orthogonal to $\bar{x}$. We apply the Ihâmessou-Pythagoras theorem to get

$$\|x^\star - \bar{x}\|^2 = (1 - \|x^\star\|\cos\theta)^2 + \|x^\star\|^2 \sin^2\theta.$$

Since $\cos\theta = \sqrt{1-\sin^2\theta}$, we have for

$$1 + \epsilon_0 \geq \sqrt{1+\epsilon_0} \geq \|x^\star\|\cos\theta \geq \sqrt{1 - \epsilon_0 - \frac{\epsilon_0^2}{1-\epsilon_0}} \geq 1 - \epsilon_0.$$

where we used that $\epsilon_0 < 1/3$ in the last inequality. We then get that

$$(1 - \|x^\star\|\cos\theta)^2 \leq \epsilon_0^2.$$

In turn

$$\|x^\star - \bar{x}\|^2 \leq \epsilon_0^2 + \frac{\epsilon_0^2(1+\epsilon_0)}{(1-\epsilon_0)^2} \leq 4\epsilon_0^2$$

for $\epsilon_0 < 1/3$. We also know that

$$\|x^\star - \bar{x}\| \leq 2 + \epsilon_0 \leq 7/3$$

for $\epsilon_0 < 1/3$. We therefore get that

$$\mathrm{dist}(x^\star, \overline{\mathcal{X}}) \leq 8\min(\bar{\epsilon}, 1) \leq 8\bar{\epsilon},$$

where the last inequality is a consequence of Assumption 4.3.1. □

4.6.3 Relative smoothness

Lemma 4.6.4. *Fix $\varrho \in]0,1[$, if the event $\mathcal{E}_{\text{conH}}$ holds true then,*

$$\forall x, z \in \mathbb{R}^n, \quad D_f(x,z) \leq \left(3 + \widetilde{\epsilon} + \varrho \max\left(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1\right)\right) D_\psi(x,z). \tag{4.6.8}$$

Proof. Let fix $\varrho \in]0,1[$, for any $u \in \mathbb{R}^n$, we have

$$\begin{aligned}
\nabla^2 f(u) &\preceq \mathbb{E}\left(\nabla^2 f(u)\right) + \varrho(\|u\|^2 + \|\bar{x}\|^2/3 + \|\epsilon\|_\infty)\mathrm{Id}, \\
&\preceq 3\left(2uu^\top + \|u\|^2 \mathrm{Id}\right) - 2\bar{x}\bar{x}^\top - \|\bar{x}\|^2 \mathrm{Id} - \widetilde{\epsilon}\mathrm{Id} + \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)\left(\|u\|^2 + 1\right)\mathrm{Id}, \\
&\preceq 3\left(2uu^\top + (\|u\|^2 + 1)\mathrm{Id}\right) + \widetilde{\epsilon}\mathrm{Id} + \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)\nabla^2\psi(x), \\
&\preceq 3\nabla^2\psi(u) + \widetilde{\epsilon}\nabla^2\psi(u) + \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)\nabla^2\psi(x), \\
&\preceq \left(3 + \widetilde{\epsilon} + \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)\right)\nabla^2\psi(u),
\end{aligned} \tag{4.6.9}$$

We conclude by applying Lemma 2.3.8 in the segment $[x,z]$. □

4.6.4 Local relative strong convexity

Lemma 4.6.5. *Fix $\lambda \in \left]\frac{1}{9\sqrt{2}}, 1\right[$ and for $\varrho \in \left]0, \frac{\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon}}{2\max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)}\right[$. If the event $\mathcal{E}_{\text{conH}}$ holds true, then for any $x, z \in B(\bar{x}, \rho)$ or $x, z \in B(-\bar{x}, \rho)$, we have*

$$D_f(x,z) \geq \left(\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon} - \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)\right) D_\psi(x,z), \tag{4.6.10}$$

where $\rho = \frac{1-\lambda}{\sqrt{3}}\|\bar{x}\|$.

Proof. For any $u \in \mathbb{R}^n$, we have

$$\begin{aligned}
\nabla^2 f(u) &\succeq \mathbb{E}\left(\nabla^2 f(u)\right) - \varrho\left(\|u\|^2 + \|\bar{x}\|^2/3 + \|\epsilon\|_\infty\right)\mathrm{Id}, \\
&\succeq 6uu^\top + 3\|u\|^2 - 2\bar{x}\bar{x}^\top - \|\bar{x}\|^2 \mathrm{Id} - \widetilde{\epsilon}\mathrm{Id} - \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)(\|u\|^2 + 1).
\end{aligned}$$

We then obtain, for any $v \in \mathbb{S}^{n-1}$

$$v^\top\left(\nabla^2 f(u) + \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)\nabla^2\psi(u)\right)v \geq 3\left(2\left(v^\top u\right)^2 + \|u\|^2\right) - \left(2\left(v^\top \bar{x}\right)^2 + \|\bar{x}\|^2\right) - \widetilde{\epsilon}\mathrm{Id}.$$

Let $\rho > 0$ be small enough, to be made precise later. Thus for any $u = \pm\bar{x} + \rho v$ we get

$$\begin{aligned}
v^\top\nabla^2 f(u)v + \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)v^\top\nabla^2\psi(u)v \\
&\geq 6\left(v^\top\bar{x}\right)^2 + 6\rho^2 \pm 12\rho v^\top\bar{x} + 3\|\bar{x}\|^2 \pm 6\rho v^\top\bar{x} + 3\rho^2 - 2\left(v^\top\bar{x}\right)^2 - \|\bar{x}\|^2 - \widetilde{\epsilon} \\
&= 4\left(v^\top\bar{x}\right)^2 + 9\rho^2 \pm 18\rho v^\top\bar{x} + 2\|\bar{x}\|^2 - \widetilde{\epsilon}.
\end{aligned} \tag{4.6.11}$$

For the entropy ψ, we also have

$$v^\top\nabla^2\psi(u)v = \|u\|^2 + 1 + 2\left(v^\top u\right)^2 = 2\left(v^\top\bar{x}\right)^2 + 3\rho^2 + \pm 6\rho v^\top\bar{x} + \|\bar{x}\|^2 + 1.$$

At this step, the proof becomes very similar to the noiseless phase retrieval (see [83, Lemma B.6]). Indeed, let us observe that we showed that for any vector $u = \pm\bar{x} + \rho v$ with $\rho = \frac{1-\lambda}{\sqrt{3}}\|\bar{x}\|$ we have,

$$4\left(v^\top\bar{x}\right)^2 + 9\rho^2 \pm 18\rho v^\top\bar{x} + 2\|\bar{x}\|^2 \geq \lambda \min(\|\bar{x}\|^2, 1)\left(2\left(v^\top\bar{x}\right)^2 + 3\rho^2 + \pm 6\rho v^\top\bar{x} + \|\bar{x}\|^2 + 1\right)$$

By replacing this result in (4.6.11), we get

$$\begin{aligned}
&v^\top\nabla^2 f(u)v + \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)v^\top\nabla^2\psi(u)v \\
&\geq \lambda \min(\|\bar{x}\|^2, 1)\left(2\left(v^\top\bar{x}\right)^2 + 3\rho^2 + \pm 6\rho v^\top\bar{x} + \|\bar{x}\|^2 + 1\right) - \widetilde{\epsilon} \\
&\geq \left(\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon}\right)v^\top\nabla^2\psi(u)v
\end{aligned}$$

Finally, we have that

$$v^\top \left(\nabla^2 f(x) - \left(\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon} - \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1) \right) \nabla^2 \psi(x) \right) v \geq 0$$

for all $v \in \mathbb{S}^{n-1}$ and $\rho \leq \frac{1-\lambda}{\sqrt{3}} \|\bar{x}\|$. To conclude the proof, let us observe that with the prescribed bound on ϱ, we have

$$\sigma = \lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon} - \varrho \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1) > \frac{\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon}}{2} > 0, \tag{4.6.12}$$

where we used Assumption 4.3.1 on the noise. Therefore, (4.6.10) follows simply by invoking Lemma 2.3.8. □

We have the following corollary which gives a condition on the coefficient of the signal-to-noise ratio c_s ensuring that the neighborhood of strong convexity ρ is greater than the noise.

Corollary 4.6.6. *For any fixed* $\lambda \in \left]\frac{1}{9\sqrt{2}}, 1\right[$ *and* $\varrho \in \left]0, \frac{\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon}}{2 \max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)}\right[$, *if Assumption 4.3.1 is satisfied then* $r^2 = \rho^2 - \frac{4\|\epsilon\|^2}{m\sigma} > 0$, *where* $\rho = \frac{1-\lambda}{\sqrt{3}} \|\bar{x}\|$.

Proof. To have the desired result, it suffices that $\rho^2 - \frac{4\|\epsilon\|^2}{m\sigma} > 0$ *i.e.* $\rho > 2\frac{\|\epsilon\|_\infty}{\sqrt{\sigma}}$. From (4.6.12) we have,

$$\sqrt{\sigma} > \frac{\sqrt{\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon}}}{\sqrt{2}}$$

thus,

$$\frac{2\|\epsilon\|_\infty}{\sqrt{\sigma}} \leq \frac{2\sqrt{2}\|\epsilon\|_\infty}{\sqrt{\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon}}} \leq \frac{2\sqrt{2} c_s \min(\|\bar{x}\|^2, 1)}{\sqrt{\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon}}}.$$

Therefore it suffices to show that

$$\rho > \frac{2\sqrt{2} c_s \min(\|\bar{x}\|^2, 1)}{\sqrt{\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon}}}.$$

Replacing now ρ by its expression, we get that c_s should satisfy

$$(1-\lambda)\|\bar{x}\| \sqrt{\lambda \min(\|\bar{x}\|^2, 1) - \widetilde{\epsilon}} > 2\sqrt{6} c_s \min(\|\bar{x}\|^2, 1)$$

which holds thanks to Assumption 4.3.1. □

Remark 4.6.7. This result estimates the maximum signal-to-noise ratio for which we ensure that the neighborhood of strong convexity around the true vectors is well-defined. Let us notice that a more practical upper bound is

$$c_s < \frac{(1-\lambda)\sqrt{\lambda}}{2\sqrt{6}} \leq \frac{1}{9\sqrt{2}}. \tag{4.6.13}$$

Indeed, it is a simple maximization of the function $\lambda \mapsto \frac{(1-\lambda)\sqrt{\lambda}}{2\sqrt{6}}$ over $\left]\frac{1}{9\sqrt{2}}, 1\right[$.

4.6.5 Spectral initialization

We now show that the initial guess x_0 generated by spectral initialization (Algorithm 4) belongs to a small f-attentive neighborhood of $\overline{\mathcal{X}}$.

Lemma 4.6.8. *Fix* $\varrho \in]0, 1[$ *and assume that for* $r \in [m]$, *we have* $\epsilon[r] \leq |a_r^\top \bar{x}|^2$. *If the number of samples obeys* $m \geq C(\varrho) n \log n$ *for some sufficiently large constant* $C(\varrho) > 0$ *and* (4.6.2) *holds true then* x_0 *satisfies:*

(i) $\operatorname{dist}(x_0, \overline{\mathcal{X}}) \leq \eta_1(\varrho)\,\|\bar{x}\|$, *where*

$$\begin{aligned}\eta_1 \colon]0,1[&\to]0,1[\\ \varrho &\mapsto \left(\sqrt{2-2\sqrt{1-\varrho(1+c_s)}} + \frac{\varrho(1+c_s)}{2}\right),\end{aligned} \tag{4.6.14}$$

which is an increasing function.

(ii) Moreover, we have

$$f(x_0) \leq f(\bar{x}) + \left((1+\varrho)c_s\,\|\bar{x}\| + L\frac{\Theta(\eta_1(\varrho)\,\|\bar{x}\|)}{2}\eta_1(\varrho)\right)\eta_1(\varrho)\,\|\bar{x}\|^2, \tag{4.6.15}$$

with $L = 3 + \widetilde{\epsilon} + \varrho\max(\|\bar{x}\|^2/3 + \|\epsilon\|_\infty, 1)$.

(iii) Besides, for $\lambda \in]0,1[$, *if*

$$\varrho \leq \eta_1^{-1}\left(\frac{1-\lambda}{\sqrt{3\left(6\left(1+\frac{(1-\lambda)^2}{3}-\frac{4\|\epsilon\|^2}{m\sigma\|\bar{x}\|^2}\right)+1\right)}}\frac{1}{\max\left(\|\bar{x}\|,1\right)}\right), \tag{4.6.16}$$

then we have $x_0 \in B\left(\overline{\mathcal{X}}, \frac{r}{\max\left(\sqrt{\Theta(r)},1\right)}\right)$ *where* $r^2 = \frac{(1-\lambda)^2}{3}\|\bar{x}\|^2 - \frac{4\|\epsilon\|^2}{m\sigma}$.

Proof. (i) Denote the matrix

$$Y = \frac{1}{m}\sum_{r=1}^{m} y[r]a_r a_r^\top = \frac{1}{m}\sum_{r=1}^{m}(|a_r^\top\bar{x}|^2 + \epsilon[r])a_r a_r^\top.$$

We have *w.h.p*

$$\|Y - \mathbb{E}(Y)\| \leq \varrho(\|\bar{x}\|^2 + \|\epsilon\|_\infty) \leq \varrho(1+c_s)\,\|\bar{x}\|^2.$$

Let $\widetilde{x}$ be the eigenvector associated with the largest eigenvalue $\widetilde{\lambda}$ of Y such that $\|\widetilde{x}\| = \|\bar{x}\|$ (obviously $\widetilde{\lambda}$ is nonnegative since Y is semidefinite positive). Then,

$$\begin{aligned}\varrho(1+c_s)\,\|\bar{x}\|^4 &\geq \left|\widetilde{x}^\top\left(Y - 2\bar{x}\bar{x}^\top - \|\bar{x}\|^2\operatorname{Id} - \widetilde{\epsilon}\operatorname{Id}\right)\widetilde{x}\right| \\ &= \left|\widetilde{\lambda}\,\|\bar{x}\|^2 - 2(\widetilde{x}^\top\bar{x})^2 - \|\bar{x}\|^4 - \widetilde{\epsilon}\,\|\bar{x}\|^2\right|\end{aligned}$$

Hence

$$2(\widetilde{x}^\top\bar{x})^2 \geq \widetilde{\lambda}\,\|\bar{x}\|^2 - \|\bar{x}\|^4 - \widetilde{\epsilon}\,\|\bar{x}\|^2 - \varrho(1+c_s)\,\|\bar{x}\|^4.$$

Moreover, since $\widetilde{\lambda}$ is the largest eigenvalue of Y, applying the concentration inequality at $\bar{x}$ *w.h.p*

$$\begin{aligned}\widetilde{\lambda}\,\|\bar{x}\|^2 \geq \bar{x}^\top Y\bar{x} &\geq \bar{x}^\top\left(2\bar{x}\bar{x}^\top + \|\bar{x}\|^2\operatorname{Id} - \widetilde{\epsilon}\operatorname{Id}\right)\bar{x} - \varrho(1+c_s)\,\|\bar{x}\|^4 \\ &= 3\,\|\bar{x}\|^4 + \widetilde{\epsilon}\,\|\bar{x}\|^2 - \varrho(1+c_s)\,\|\bar{x}\|^4.\end{aligned}$$

Merging the last two inequalities, we get

$$\begin{aligned}2(\widetilde{x}^\top\bar{x})^2 &\geq 3\,\|\bar{x}\|^4 + \widetilde{\epsilon}\,\|\bar{x}\|^2 - \varrho(1+c_s)\,\|\bar{x}\|^4 - \|\bar{x}\|^4 - \widetilde{\epsilon}\,\|\bar{x}\|^2 - \varrho(1+c_s)\,\|\bar{x}\|^4 \\ &= 2\,\|\bar{x}\|^4 - 2\varrho(1+c_s)\,\|\bar{x}\|^4.\end{aligned}$$

Which implies that

$$\operatorname{dist}(\widetilde{x}, \overline{\mathcal{X}}) \leq \sqrt{2-2\sqrt{1-\varrho(1+c_s)}}\,\|\bar{x}\|.$$

By definition of x_0 in Algorithm 4, $x_0 = \sqrt{\frac{1}{m}\sum_r y[r]}\frac{\widetilde{x}}{\|\widetilde{x}\|} = \sqrt{\frac{1}{m}\sum_r(|a_r^\top\bar{x}|^2 + \epsilon[r])}\frac{\widetilde{x}}{\|\widetilde{x}\|}$, and thus *w.h.p*

$$\|x_0 - \widetilde{x}\| = \left|\sqrt{\frac{m^{-1}\sum_r y[r]}{\|\bar{x}\|^2}} - 1\right|\|\bar{x}\| = \left|\sqrt{\frac{m^{-1}\sum_r(|a_r^\top\bar{x}|^2 + \epsilon[r])}{\|\bar{x}\|^2}} - 1\right|\|\bar{x}\| \leq \frac{\varrho(1+c_s)\,\|\bar{x}\|}{2},$$

it comes out that,

$$\operatorname{dist}(x_0, \overline{\mathcal{X}}) \leq \operatorname{dist}(\widetilde{x}, \overline{\mathcal{X}}) + \|x_0 - \widetilde{x}\| \leq \left(\sqrt{2 - 2\sqrt{1 - \varrho(1+c_s)}} + \frac{\varrho(1+c_s)}{2} \right) \|\bar{x}\| .$$

(ii) Under our sampling complexity bound, event $\mathcal{E}_{\mathrm{conH}}$ defined by (4.3.3) holds true *w.h.p* . It then follows from Lemma 4.6.4 applied at $\bar{x}$ and x_0, that

$$D_f(x_0, \bar{x}) \leq L D_\psi(x_0, \bar{x}).$$

The latter implies that

$$f(x_0) \leq f(\bar{x}) + \langle \nabla f(\bar{x}); x_0 - \bar{x} \rangle + L D_\psi(x_0, \bar{x}),$$
$$f(x_0) \leq f(\bar{x}) + \|\nabla f(\bar{x})\| \|x_0 - \bar{x}\| + L \frac{\Theta(\eta_1(\varrho)\|\bar{x}\|)}{2} \|x_0 - \bar{x}\|^2 ,$$

Since $\nabla f(\bar{x}) = \frac{1}{m} \sum_{r=1}^{m} \epsilon[r] a_r a_r^\top \bar{x}$, we obtain from (4.6.3) that

$$\|\nabla f(\bar{x})\| \leq \|\epsilon\|_\infty \left\| \frac{1}{m} \sum_{r=1}^{m} a_r a_r^\top \right\| \|\bar{x}\| \leq (1+\varrho) \|\epsilon\|_\infty \|\bar{x}\| .$$

Combining the two last inequality yields to

$$f(x_0) \leq f(\bar{x}) + \left((1+\varrho) c_s \|\bar{x}\| + L \frac{\Theta(\eta_1(\varrho)\|\bar{x}\|)}{2} \eta_1(\varrho) \right) \eta_1(\varrho) \|\bar{x}\|^2 .$$

(iii) In view of (i), it is sufficient to show that $\eta_1(\varrho)\|\bar{x}\| \leq \frac{r}{\max\left(\sqrt{\Theta(r)}, 1\right)}$. Since from Proposition 2.3.5(iv) (see also Remark 3.2.8) we have

$$\Theta(r) \leq 6(\|\bar{x}\|^2 + r^2) + 1 \leq \left(6 \left(1 + \frac{(1-\lambda)^2}{3} - \frac{4\|\epsilon\|^2}{m\sigma\|\bar{x}\|^2} \right) + 1 \right) \max\left(\|\bar{x}\|^2 , 1 \right),$$

and η_1 is an increasing function, we conclude.

□

4.7 Landscape of the Noise-Aware Objective with Gaussian Measurements

4.7.1 Warm up: Critical points of $\mathbb{E}(f)$

We start by studying and characterizing the set of critical points of $\mathbb{E}(f)$. This can be seen as the asymptotic behavior of the critical points of f when the number of measurements m grows to $+\infty$.

Proposition 4.7.1. *We have*

$$\operatorname{crit}(\mathbb{E}(f)) = \{0\} \cup \overline{\mathcal{X}}_\epsilon \cup \left\{ x \in \mathbb{R}^n : \bar{x}^\top x = 0, \|x\|^2 = \frac{1}{3}\left(\|\bar{x}\|^2 + \widetilde{\epsilon} \right) \right\},$$

where $\overline{\mathcal{X}}_\epsilon \overset{\text{def}}{=} \left\{ \pm \bar{x} \sqrt{1 + \frac{\widetilde{\epsilon}}{3\|\bar{x}\|^2}} \right\}$. *Those sets are respectively, the local maximizer, the set of global minimizes, and strict saddle points of* $\mathbb{E}(f)$.

Before proving this result, we the closed form expressions of the expectation of f and its derivatives.

Lemma 4.7.2. *For all* $x \in \mathbb{R}^n$, *we have:*

$$\mathbb{E}(f(x)) = \frac{3}{4}\left(\|x\|^4 + \|\bar{x}\|^4 \right) - \frac{1}{2}\|\bar{x}\|^2 \|x\|^2 - (|\bar{x}^\top x|^2 + \frac{\|\epsilon\|^2}{4m} - \frac{\widetilde{\epsilon}(\|x\|^2 - \|\bar{x}\|^2)}{2},$$
$$\nabla \mathbb{E}(f(x)) = 3\|x\|^2 x - 2\bar{x}(\bar{x}^\top x) - \|\bar{x}\|^2 x - \widetilde{\epsilon} x,$$
$$\nabla^2 \mathbb{E}(f(x)) = 3\left(2xx^\top + \|x\|^2 \mathrm{Id} \right) - 2\bar{x}\bar{x}^\top - \|\bar{x}\|^2 \mathrm{Id} - \widetilde{\epsilon}\mathrm{Id}.$$

Proof. By linearity of the expectation, we have $\mathbb{E}\left(f(x)\right) = \mathbb{E}\left(f_{\mathrm{NL}}(x)\right) + \mathbb{E}\left(f_{\mathrm{Ny}}(x)\right)$. Linearity again yields

$$\begin{aligned}\mathbb{E}\left(f_{\mathrm{Ny}}(x)\right) &= -\frac{1}{2m}\sum_{r=1}^{m}\epsilon[r]\left(x^\top\mathbb{E}\left(a_r a_r^\top\right)x - \bar{x}^\top\mathbb{E}\left(a_r a_r^\top\right)\bar{x}\right) + \frac{\|\epsilon\|^2}{4m},\\ &= -\frac{1}{2m}\sum_{r=1}^{m}\epsilon[r]\left(\|x\|^2 - \|\bar{x}\|^2\right) + \frac{\|\epsilon\|^2}{4m}. \end{aligned} \tag{4.7.1}$$

We also have

$$\mathbb{E}\left(f_{\mathrm{NL}}(x)\right) = \frac{1}{4m}\sum_{r=1}^{m}\mathbb{E}\left(|a_r{}^\top x|^4\right) + \frac{1}{4m}\sum_{r=1}^{m}\mathbb{E}\left(|a_r{}^\top \bar{x}|^4\right) - \frac{1}{2m}\sum_{r=1}^{m}|a_r{}^\top x|^2|a_r{}^\top\bar{x}|^2.$$

From [83, Lemma B.1], we know that

$$\forall x\in\mathbb{R}^n,\quad \mathbb{E}\left(\sum_{r=1}^{m}|a_r{}^\top x|^2 a_r a_r^\top\right) = 2xx^\top + \|x\|^2\,\mathrm{Id}.$$

Therefore we have

$$\mathbb{E}\left(\frac{1}{4m}\sum_{r=1}^{m}|a_r{}^\top x|^4\right) = x^\top\left(\mathbb{E}\left(\frac{1}{4m}\sum_{r=1}^{m}|a_r{}^\top x|^2 a_r a_r^\top\right)\right)x = \frac{1}{4}x^\top\left(2xx^\top + \|x\|^2\right)x = \frac{3}{4}\|x\|^4,$$

and

$$\mathbb{E}\left(\frac{1}{2m}\sum_{r=1}^{m}|a_r{}^\top x|^2|a_r{}^\top\bar{x}|^2\right) = x^\top\left(\bar{x}\bar{x}^\top + \frac{1}{2}\|\bar{x}\|^2\right)x = |\bar{x}^\top x|^2 + \frac{1}{2}\|\bar{x}\|^2\|x\|^2.$$

Whence we have

$$\mathbb{E}\left(f_{\mathrm{NL}}(x)\right) = \frac{3}{4}\left(\|x\|^4 + \|\bar{x}\|^4\right) - \frac{1}{2}\|\bar{x}\|^2\|x\|^2 - |\bar{x}^\top x|^2. \tag{4.7.2}$$

The claim follows simply by summing (4.7.1) and (4.7.2).

We deduce the gradient and the Hessian by straightforward derivation of $\mathbb{E}\left(f\right)$. □

Proof. (Proposition 4.7.1)

- **The origin :** Let us observe that $\nabla\mathbb{E}\left(f(0)\right) = 0$, it follows that the value of the Hessian at zero satisfies

$$\nabla^2\mathbb{E}\left(f(0)\right) = -\|\bar{x}\|^2\,\mathrm{Id} - 2\bar{x}\bar{x}^\top - \widetilde{\epsilon}\mathrm{Id} \preceq -(\|\bar{x}\|^2 + \widetilde{\epsilon})\mathrm{Id} - 2\bar{x}\bar{x}^\top \prec 0,$$

  where we have used that $\widetilde{\epsilon}$ is non-negative (see Assumption 4.3.1). It follows that 0 is a local maximizer of $\mathbb{E}\left(f\right)$.

- **Subspaces with no critical points:**
  1. For any point in $\left\{x\in\mathbb{R}^n : 0 < \|x\|^2 < \frac{\|\bar{x}\|^2+\widetilde{\epsilon}}{3}\right\}$, we have

$$\langle x, \nabla\mathbb{E}\left(f(x)\right)\rangle = \left(3\|x\|^2 - \|\bar{x}\|^2 - \widetilde{\epsilon}\right)\|x\|^2 - 2|\bar{x}^\top x|^2 \le \left(3\|x\|^2 - \|\bar{x}\|^2 - \widetilde{\epsilon}\right)\|x\|^2 < 0.$$

     We deduce that, $\langle x, \nabla\mathbb{E}\left(f(x)\right)\rangle < 0$ which implies that $\|\nabla\mathbb{E}\left(f(x)\right)\|$ is bounded away from zero on this region, and thus that there are no critical points there.

  2. Consider a point in $\left\{x\in\mathbb{R}^n : \frac{\|\bar{x}\|^2+\widetilde{\epsilon}}{3} < \|x\|^2 < \|\bar{x}\|^2 + \frac{\widetilde{\epsilon}}{3}\right\}$ and recall that it is a critical point if and only if

$$\left(3\|x\|^2 - \|\bar{x}\|^2 - \widetilde{\epsilon}\right)x = 2(\bar{x}\bar{x}^\top)x. \tag{4.7.3}$$

     Combining Assumption 4.3.1 and the fact that $\|x\|^2 > \frac{\|\bar{x}\|^2+\widetilde{\epsilon}}{3}$, we get that $\left(3\|x\|^2 - \|\bar{x}\|^2 - \widetilde{\epsilon}\right)/2$ is the positive eigenvalue of the rank-one matrix matrix $\bar{x}\bar{x}^\top$. This is equivalent to $3\|x\|^2 - \|\bar{x}\|^2 - \widetilde{\epsilon} = 2\|\bar{x}\|^2$, *i.e.*, $\|x\|^2 = \|\bar{x}\|^2 + \frac{\widetilde{\epsilon}}{3}$ which contradicts the definition of this region, showing again that there are no critical points there.

3. For a point in the region $\left\{x \in \mathbb{R}^n : \|x\|^2 > \|\bar{x}\|^2 + \frac{\widetilde{\epsilon}}{3}\right\}$, we have the lower bound

$$\begin{aligned}\langle x, \nabla \mathbb{E}\left(f(x)\right)\rangle &= \left(3\|x\|^2 - \|\bar{x}\|^2 - \widetilde{\epsilon}\right)\|x\|^2 - 2|\bar{x}^\top x|^2, \\ &\geq \left(3\|x\|^2 - 3\|\bar{x}\|^2 - \widetilde{\epsilon}\right)\|x\|^2 > 0.\end{aligned}$$

Hence, $\|\nabla \mathbb{E}\left(f(x)\right)\|$ is bounded away from zero on this region yielding the same conclusion.

- **Strict saddle points:** A point in the sphere $\left\{x \in \mathbb{R}^n : \|x\|^2 = \frac{\|\bar{x}\|^2 + \widetilde{\epsilon}}{3}\right\}$ is a critical point if it is orthogonal to the true vector $\bar{x}$. Indeed we have,

$$\nabla \mathbb{E}\left(f(x)\right) = 0 \iff \left(3\|x\|^2 - \|\bar{x}\|^2 - \widetilde{\epsilon}\right)x = 2\bar{x}\bar{x}^\top x \iff \bar{x}^\top x = 0.$$

Besides, for any $v \in \mathbb{R}^n$ we have

$$\begin{aligned}\left\langle v, \nabla^2 \mathbb{E}\left(f(x)\right) v\right\rangle &= 6|v^\top x|^2 + 3\|x\|^2\|v\|^2 - 2|v^\top \bar{x}|^2 - \|\bar{x}\|^2\|v\|^2 - \widetilde{\epsilon}\|v\|^2 \\ &= 6|v^\top x|^2 - 2|v^\top \bar{x}|^2 + \|v\|^2\left(3\|x\|^2 - \|\bar{x}\|^2 - \widetilde{\epsilon}\right) \\ &= 6|v^\top x|^2 - 2|v^\top \bar{x}|^2.\end{aligned}$$

In the direction $v = x$, we deduce that

$$\left\langle x, \nabla^2 \mathbb{E}\left(f(x)\right) x\right\rangle = 6\|x\|^4 > 0,$$

and in the direction $v = \bar{x}$ we have

$$\left\langle \bar{x}, \nabla^2 \mathbb{E}\left(f(x)\right) \bar{x}\right\rangle = -2\|\bar{x}\|^4 < 0,$$

where we have used orthogonality of x and $\bar{x}$. These facts show that the critical points in this region, *i.e.* points orthogonal to $\bar{x}$, are strict saddle points of $\mathbb{E}\left(f\right)$.

- **Global minimizers:** In view of the above, local/global minimizers can only occur on the sphere $\left\{x \in \mathbb{R}^n : \|x\|^2 = \|\bar{x}\|^2 + \frac{\widetilde{\epsilon}}{3}\right\}$. Any point in his set is a critical point of $\mathbb{E}\left(f\right)$ if and only if

$$\nabla \mathbb{E}\left(f(x)\right) = 0 \iff \left(3\|x\|^2 - \|\bar{x}\|^2 - \widetilde{\epsilon}\right)x = 2\bar{x}\bar{x}^\top x \iff \left(\|\bar{x}\|^2 \operatorname{Id} - \bar{x}\bar{x}^\top\right)x = 0.$$

Therefore, x is a critical point on this region if and only if x is an eigenvector of $\bar{x}\bar{x}^\top$, that is $x \in \operatorname{span}(\bar{x})$, or equivalently, $\exists \beta \in \mathbb{R}$ such that $x = \beta\bar{x}$ with

$$\|x\|^2 = \beta^2\|\bar{x}\|^2 = \|\bar{x}\|^2 + \frac{\widetilde{\epsilon}}{3} \iff \beta = \pm\sqrt{1 + \frac{\widetilde{\epsilon}}{3\|\bar{x}\|^2}}.$$

The set of critical points in this region is reduced to $\overline{\mathcal{X}}_\epsilon \overset{\text{def}}{=} \left\{\pm\bar{x}\sqrt{1 + \frac{\widetilde{\epsilon}}{3\|\bar{x}\|^2}}\right\}$. For $x \in \overline{\mathcal{X}}_\epsilon$ we have

$$\begin{aligned}\nabla^2 \mathbb{E}\left(f(x)\right) &= 6xx^\top - 2\bar{x}\bar{x}^\top + \left(3\|x\|^2 - \|\bar{x}\|^2 - \widetilde{\epsilon}\right)\operatorname{Id} \\ &= (6\beta^2 - 2)\bar{x}\bar{x}^\top + 2\|\bar{x}\|^2 \operatorname{Id} \succeq 2\|\bar{x}\|^2 \operatorname{Id}.\end{aligned}$$

Indeed, we have

$$6\beta^2 - 2 = 4 + \frac{2\widetilde{\epsilon}}{\|\bar{x}\|^2} \geq 4 > 0,$$

where we use again the non-negativity of $\widetilde{\epsilon}$ in Assumption 4.3.1. We conclude that $\overline{\mathcal{X}}_\epsilon$ is the set of global minimizers of $\mathbb{E}\left(f\right)$.

□

4.7.2 Main result: Critical points of f

In this section, we study the landscape of the objective function f for the Gaussian measurement model. Our main result hereafter characterizes the set of critical points of f for m large enough.

Theorem 4.7.3. ***(Critical points of f)*** *Fix $\lambda \in \left]\frac{1}{9\sqrt{2}}, 1\right[$. Let us assume that the noise vector satisfies Assumption 4.3.1. If $m \gtrsim n\log(n)^3$, then*

$$\operatorname{crit}(f) = \operatorname{Argmin}(f) \cup \operatorname{strisad}(f) \tag{4.7.4}$$

where $\operatorname{Argmin}(f) = \{\pm x^\star\}$. *This holds with probability of at least* $1 - \frac{c}{m}$ *where c is a positive numerical constant.*

Remark 4.7.4.

- In [168, Theorem 2.2], the authors study the geometry of f in the noiseless case here coined as f_{NL}. We aim with our result to extend it to the noisy case with small enough noise (see Assumption 4.3.1).
- This result shows that when the number of measurements m is sufficiently large and the noise ϵ is very small compared to the true vector which is entailed by a large SNR, then the set of critical points of the objective function f is reduced to the set of global minimizers $\operatorname{Argmin}(f)$ and the set of strict saddle points $\operatorname{strisad}(f)$. The strict saddle avoidance of mirror descent will then imply that the sequence provided by mirror descent will always converge to global minimizers of the function f.

We recall the radius $\rho = \frac{1-\lambda}{\sqrt{3}}\|\bar{x}\|$ defined in Lemma 4.6.5. To prove Theorem 4.7.3, we consider the following regions of $\mathbb{R}^n$ which are helpful to characterize the landscape of f:

$$\mathcal{R}_1 = \left\{x \in \mathbb{R}^n : \left\langle \bar{x}, \mathbb{E}\left(\nabla^2 f(x)\right)\bar{x}\right\rangle \leq -\frac{1}{100}\|x\|^2\|\bar{x}\|^2 - \frac{1}{50}\|\bar{x}\|^4\right\}, \tag{4.7.5}$$

$$\mathcal{R}_3 = \left\{x \in \mathbb{R}^n : \operatorname{dist}(x, \overline{\mathcal{X}}) \leq \rho\right\}, \tag{4.7.6}$$

$$\mathcal{R}_2 = (\mathcal{R}_1 \cup \mathcal{R}_3)^c. \tag{4.7.7}$$

We also define specific regions $\mathcal{R}_2^x$ and $\mathcal{R}_2^h$,

$$\mathcal{R}_2^x = \left\{x \in \mathbb{R}^n : \langle x, \mathbb{E}\left(\nabla f(x)\right)\rangle \geq \frac{1}{500}\|x\|^2\|\bar{x}\|^2 + \frac{1}{100}\|x\|^4\right\},$$

$$\mathcal{R}_2^h = \left\{x \in \mathbb{R}^n : \langle d_x, \mathbb{E}\left(\nabla f(x)\right)\rangle \geq \frac{1}{250}\|x\|\|d_x\|\|\bar{x}\|^2, \frac{11}{20}\|\bar{x}\| \leq \|x\| \leq \|\bar{x}\|, \operatorname{dist}(x, \overline{\mathcal{X}}) \geq \frac{\|\bar{x}\|}{3}\right\}.$$

where $d_x = \frac{\pm\bar{x}-x}{\|\pm\bar{x}-x\|}$ if $x \neq \pm\bar{x}$ and any vector on the unit sphere otherwise.

Let us observe that these regions are similar to those defined in [168] replacing f by f_{NL}. Indeed, the idea behind our assumptions is the fact that small noise will introduce small perturbations in the function f, and therefore under our assumption of small noise, the latter has benign influence on the landscape of f (see Figure 4.5). Mainly, in the region $\mathcal{R}_1$ the function f still has negative curvature. In the region $\mathcal{R}_2$, f has a large gradient and in $\mathcal{R}_3$ relative strong convexity with respect to our chosen entropy ψ. It is important to observe that in the noisy case, it is not true that the true vectors $\overline{\mathcal{X}}$ are critical points of f or even $\mathbb{E}(f)$. However, we have already shown in Lemma 4.5.2 that $\pm\bar{x}$ are actually $\frac{\|\epsilon\|^2}{m}$-minimizers. Moreover, we have already given in Proposition 4.7.1 a description of the set of critical points of $\mathbb{E}(f)$, providing a hint that in the large oversampling regime, the geometry of f is close to that of f_{NL}. This result shows that the set of critical points of $\mathbb{E}(f)$ is also reduced to the set of strict saddle points with symmetric global minimizers of $\mathbb{E}(f)$. This set of minimizers, that we denoted $\overline{\mathcal{X}}_\epsilon$ (see Proposition 4.7.1), are direct perturbations of the true vectors $\overline{\mathcal{X}}$ by the noise; see also Lemma 4.6.3 which quantifies the distance of global minimizers of f to $\overline{\mathcal{X}}$ in probability.

Proof. In the following, all assertions are to be understood in high probability sense. The proof consists in invoking properly the statements of Proposition 4.7.5. In the region $\mathcal{R}_1$, Proposition 4.7.5-(i) shows that

$$\forall x \in \mathcal{R}_1, \quad \left\langle \bar{x}, \nabla^2 f(x)\bar{x}\right\rangle \leq -\frac{1}{100}(1 - c_s)\|\bar{x}\|^4, \tag{4.7.8}$$

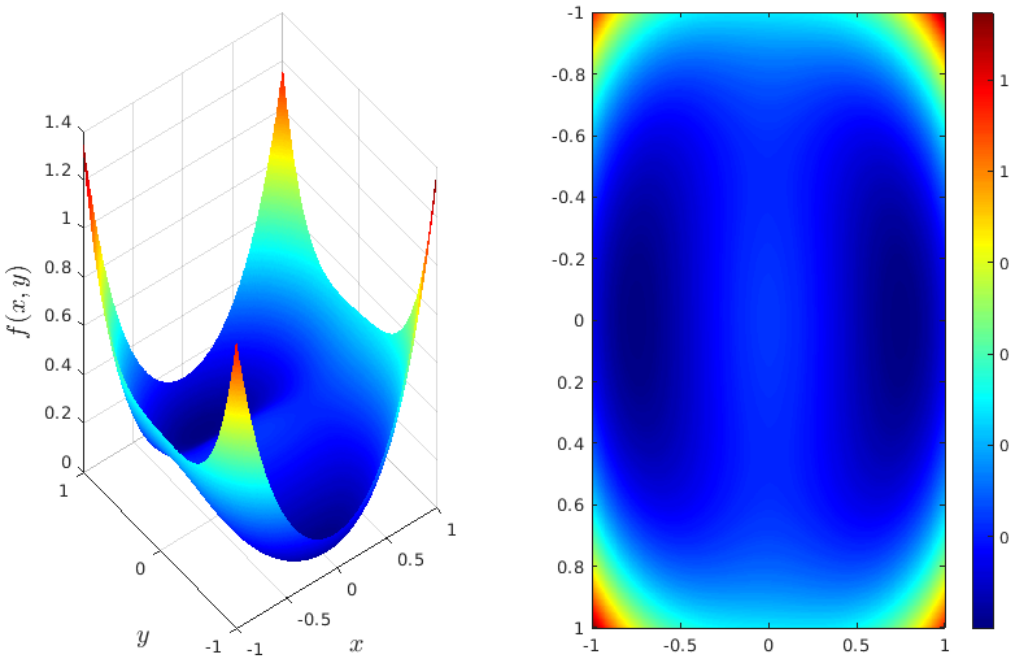


Figure 4.5: Landscape of the function f as $m \to \infty$; we have $(m, n) = (200, 2)$ and the true vectors are $[\pm 3/4, 0]$. The noise vector is generated at uniform in [-1,1] such that $\widetilde{\epsilon} \approx 5.10^{-3}$. One clearly sees that the geometry of the landscape of f is preserved and that the only minimizers of f are very close to the true vectors.

i.e. f has a negative curvature in the direction of the true vectors $\overline{\mathcal{X}}$ which means that any critical point in $\mathcal{R}_1$ is a strict saddle point for f. From Proposition 4.7.5-(iii) and (iv), we deduce that

$$\forall x \in \mathcal{R}_2^x \cup \mathcal{R}_2^h, \quad \|\nabla f(x)\| \geq \frac{1}{1000}(1 - c_s) \|x\| \|\bar{x}\|^2 . \tag{4.7.9}$$

Moreover, Proposition 4.7.5-(v) entails that $\mathcal{R}_2 \subset \mathcal{R}_2^x \cup \mathcal{R}_2^h$ which means that (4.7.9) holds true for all $x \in \mathcal{R}_2$. Thus the gradient of function f is bounded away from zero on $\mathcal{R}_2$ which means that there are no critical points in this region. Therefore, local/global minimizers of f are necessarily located in the region $\mathcal{R}_3$. It remains to show that the only critical points in the domain $\mathcal{R}_3$ are just the elements of Argmin(f)[2] which contains only two points $\pm x^\star$. This will be a consequence of σ-strong convexity of f on $\mathcal{R}_3$. In the following, since $\mathcal{R}_3 = B(\bar{x}, \rho) \cup B(-\bar{x}, \rho)$, we prove the claim only on $B(\bar{x}, \rho)$ and the same holds for the symmetric case with $-\bar{x}$. Let $x \in B(\bar{x}, \rho) \backslash \{x^\star\}$. In view of Proposition 4.7.5-(ii), we have

$$D_f(x, x^\star) = f(x) - \min f \geq \sigma D_\psi(x, x^\star) \geq \frac{\sigma}{2} \|x - x^\star\|^2 .$$

The right hand side is positive since $x \neq x^\star$, which means that f has a unique minimizer on $B(\bar{x}, \rho)$. Moreover,

$$D_f(x^\star, x) = \min f - f(x) - \langle \nabla f(x), x^\star - x \rangle \geq \sigma D_\psi(x^\star, x) \geq \frac{\sigma}{2} \|x - x^\star\|^2 ,$$

and thus

$$\langle \nabla f(x), x - x^\star \rangle \geq D_f(x^\star, x) \geq \frac{\sigma}{2} \|x - x^\star\|^2 .$$

Cauchy-Schwarz then entails

$$\|\nabla f(x)\| \geq \frac{\sigma}{2} \|x - x^\star\| > 0$$

meaning that f has no other critical point than $x^\star$ on $B(\bar{x}, \rho)$. This completes the proof. □

[2]Remember that Argmin(f) is a nonempty compact set by injectivity of A under the assumed measurement bound.

The proof of the above result heavily relies on the behaviour of f on each region. This is the subject of the next proposition.

Proposition 4.7.5. *If the number of samples obeys $m \gtrsim n\log^3(n)$ then with probability $1-\frac{c}{m}$ where c is a positive numerical constant, we have the following statements.*

(i) In the region $\mathcal{R}_1$, the objective f has a negative curvature i.e.,

$$\forall x \in \mathcal{R}_1, \quad \left\langle \bar{x}, \nabla^2 f(x)\bar{x} \right\rangle \leq -\frac{1}{100}(1-c_s)\left\|\bar{x}\right\|^4. \tag{4.7.10}$$

(ii) In $\mathcal{R}_3$, f is σ-strongly convex where $\sigma > 0$ is given in Proposition 4.6.5.

(iii) The gradient is bounded from away from zero in $\mathcal{R}_2^x$. More precisely,

$$\forall x \in \mathcal{R}_2^x, \quad \langle x, \nabla f(x)\rangle \geq \frac{1}{1000}(1-c_s)\left\|x\right\|^2\left\|\bar{x}\right\|^2. \tag{4.7.11}$$

(iv) We have

$$\forall x \in \mathcal{R}_2^h, \quad \langle d_x, \nabla f(x)\rangle \geq \frac{1}{1000}(1-c_s)\left\|\bar{x}\right\|^2\left\|x\right\|\left\|d_x\right\|. \tag{4.7.12}$$

(v) We have $\mathcal{R}_2 \subset \mathcal{R}_2^x \cup \mathcal{R}_2^h$.

Remark 4.7.6. The previous result extends the series of propositions ([168, Proposition 2.3-2.7]) to the noisy case. All the statements depend on the (inverse) signal-to-noise coefficient c_s which obviously less than 1 under our assumption. In the noiseless case, let us observe that we recover all the Propositions mentioned above.

Proof.

(i) For any $x \in \mathcal{R}_1$, we have

$$\left\langle \bar{x}, \nabla^2 f(x)\bar{x} \right\rangle = \frac{1}{m}\sum_{r=1}^m 3\left|a_r^\top x\right|^2\left|a_r^\top \bar{x}\right|^2 - \frac{1}{m}\sum_{r=1}^m \left|a_r^\top \bar{x}\right|^4 - \frac{1}{m}\sum_{r=1}^m \epsilon[r]\left|a_r^\top \bar{x}\right|^2.$$

By using similar concentration inequalities as in Lemma 4.6.2 we have the following

$$\begin{aligned}
\frac{1}{m}\sum_{r=1}^m 3\left|a_r^\top x\right|^2\left|a_r^\top \bar{x}\right|^2 &\leq \mathbb{E}\left(\frac{1}{m}\sum_{r=1}^m 3\left|a_r^\top x\right|^2\left|a_r^\top \bar{x}\right|^2\right) + \varrho\left\|\bar{x}\right\|^2\left\|x\right\|^2, \\
\frac{1}{m}\sum_{r=1}^m \left|a_r^\top \bar{x}\right|^4 &\geq \mathbb{E}\left(\frac{1}{m}\sum_{r=1}^m \left|a_r^\top \bar{x}\right|^4\right) - \varrho\left\|\bar{x}\right\|^4, \\
\frac{1}{m}\sum_{r=1}^m \epsilon[r]\left|a_r^\top x\right|^2 &\geq \mathbb{E}\left(\frac{1}{m}\sum_{r=1}^m \epsilon[r]\left|a_r^\top x\right|^2\right) - \varrho\left\|\epsilon\right\|_\infty\left\|x\right\|^2.
\end{aligned}$$

After summing, we get

$$\left\langle \bar{x}, \nabla^2 f(x)\bar{x} \right\rangle \leq \left\langle \bar{x}, \mathbb{E}\left(\nabla^2 f(x)\right)\bar{x} \right\rangle + \varrho\left\|x\right\|^2\left\|\bar{x}\right\|^2 + \varrho\left\|\bar{x}\right\|^4 + \varrho c_s\left\|\bar{x}\right\|^2\left\|x\right\|^2.$$

We choose now $\varrho = \frac{1}{100}$, and since $x \in \mathcal{R}_1$, we finally obtain that

$$\begin{aligned}
\left\langle \bar{x}, \nabla^2 f(x)\bar{x} \right\rangle &\leq -\frac{1}{100}\left\|x\right\|^2\left\|\bar{x}\right\|^2 - \frac{1}{50}\left\|\bar{x}\right\|^4 + \frac{1}{100}\left\|x\right\|^2\left\|\bar{x}\right\|^2 + \frac{1}{100}\left\|\bar{x}\right\|^4 + \frac{1}{100}c_s\left\|\bar{x}\right\|^2\left\|x\right\|^2, \\
&= -\frac{1}{100}(1-c_s)\left\|\bar{x}\right\|^4.
\end{aligned}$$

(ii) Combine Lemma 4.6.5 and 1−strong convexity of ψ (see Proposition 3.2.2).

(iii) Let $x \in \mathcal{R}_2^x$,

$$\langle x, \nabla f(x)\rangle = \frac{1}{m}\sum_{r=1}^m \left|a_r^\top x\right|^4 - \frac{1}{m}\sum_{r=1}^m \left|a_r^\top \bar{x}\right|^2\left|a_r^\top \bar{x}\right|^2 - \frac{1}{m}\sum_{r=1}^m \epsilon[r]\left|a_r^\top x\right|^2.$$

using the same concentration arguments as in the proof of Lemma 4.6.2, we get

$$\begin{aligned}\langle x, \nabla f(x)\rangle &\geq \langle x, \mathbb{E}\left(\nabla f(x)\right)\rangle - \frac{1}{100}\left\|x\right\|^4 - \frac{1}{1000}\left\|x\right\|^2\left\|\bar{x}\right\|^2 - \frac{\left\|x\right\|^2\left\|\epsilon\right\|_\infty}{1000},\\ &\geq \frac{1}{500}\left\|x\right\|^2\left\|\bar{x}\right\|^2 + \frac{1}{100}\left\|x\right\|^4 - \frac{1}{100}\left\|x\right\|^4 - \frac{1}{1000}\left\|x\right\|^2\left\|\bar{x}\right\|^2 - \frac{\left\|x\right\|^2\left\|\epsilon\right\|_\infty}{1000},\\ &= \frac{1}{1000}(1-c_s)\left\|\bar{x}\right\|^2\left\|x\right\|^2,\end{aligned}$$

where we used Assumption 4.3.1 in the last inequality.

(iv) We have,

$$\langle d_x, \nabla f(x)\rangle = \langle d_x, \nabla f_{\mathrm{NL}}(x)\rangle + \langle d_x, \nabla f_{\mathrm{Ny}}(x)\rangle.$$

Therefore by [168, Proposition 2.6], when $m \geq Cn\log(n)^3$ with a probability at least $1-\frac{c}{m}$ we have

$$\langle d_x, \nabla f_{\mathrm{NL}}(x)\rangle \geq \langle d_x, \mathbb{E}\left(\nabla f_{\mathrm{NL}}(x)\right)\rangle - \frac{1}{500}\left\|\bar{x}\right\|^2\left\|x\right\|\left\|d_x\right\|.$$

On the other hand, with similar arguments as in the proof of Lemma 4.6.2 and using again Assumption 4.3.1, we have

$$\langle d_x, \nabla f_{\mathrm{Ny}}(x)\rangle = d_x^\top\left(\frac{1}{m}\sum_{r=1}^m \epsilon[r]a_r a_r^\top\right)x \geq \langle d_x, \mathbb{E}\left(\nabla f_{\mathrm{Ny}}(x)\right)\rangle - \frac{c_s}{500}\left\|\bar{x}\right\|^2\left\|x\right\|\left\|d_x\right\|.$$

We combine now the last two inequalities to get

$$\langle d_x, \nabla f(x)\rangle \geq \langle d_x, \mathbb{E}\left(\nabla f_{\mathrm{NL}}(x)\right) + f_{\mathrm{Ny}}(x)\rangle - \frac{1}{500}\left\|\bar{x}\right\|^2\left\|x\right\|\left\|d_x\right\| - \frac{c_s}{500}\left\|\bar{x}\right\|^2\left\|x\right\|\left\|d_x\right\|.$$

Thus for any $x \in \mathcal{R}_2^h$ we have,

$$\langle d_x, \nabla f_{\mathrm{NL}}(x)\rangle \geq \frac{1}{500}(1-c_s)\left\|\bar{x}\right\|^2\left\|x\right\|\left\|d_x\right\|.$$

(v) The proof is similar to that of [168, Proposition 2.7] which consists of showing that $\mathbb{R}^n = \mathcal{R}_1 \cup \mathcal{R}_2^x \cup \mathcal{R}_2^h \cup \mathcal{R}_3$. We then get out claim by definition of $\mathcal{R}_2$ and that $\mathbb{R}^n = \mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3$. The idea is to divide the $\mathbb{R}^n$ into several overlapping regions and show that we can cover them with our good partition. To achieve this task we will use the set

$$\mathcal{R}_2^{h'} = \left\{x \in \mathbb{R}^n : \langle d_x, \mathbb{E}\left(\nabla f(x)\right)\rangle \geq \frac{1}{250}\left\|\bar{x}\right\|^2\left\|x\right\|\left\|d_x\right\|, \left\|x\right\| \leq \left\|\bar{x}\right\|\right\}.$$

- We can cover the set $\mathcal{R}_a \overset{\text{def}}{=} \left\{x \in \mathbb{R}^n : |x^\top\bar{x}| \leq \frac{1}{2}\left\|x\right\|\left\|\bar{x}\right\|\right\}$ with both $\mathcal{R}_1$ and $\mathcal{R}_2^x$. If $\left\|x\right\|^2 \leq \frac{298}{451}\left\|\bar{x}\right\|^2$

$$\begin{aligned}\left\langle \bar{x}; \nabla^2 f(x)\bar{x}\right\rangle + \frac{1}{100}\left\|\bar{x}\right\|^2\left\|x\right\|^2 &= 6(x^\top\bar{x})^2 + \frac{301}{300}\left\|x\right\|^2\left\|\bar{x}\right\|^2 - 3\left\|\bar{x}\right\|^4 - \widetilde{\epsilon}\left\|\bar{x}\right\|^2\\ &\leq \left(\frac{3}{2}+\frac{301}{100}\right)\left\|\bar{x}\right\|^2\left\|x\right\|^2 - \frac{149}{50}\left\|\bar{x}\right\|^4 - \widetilde{\epsilon}\left\|\bar{x}\right\|^2 - \frac{1}{50}\left\|\bar{x}\right\|^4\\ &\leq \frac{298}{100}\left\|\bar{x}\right\|^4 - \frac{149}{50}\left\|\bar{x}\right\|^4 - \widetilde{\epsilon}\left\|\bar{x}\right\|^2 - \frac{1}{50}\left\|\bar{x}\right\|^4\\ &\leq \frac{1}{50}\left\|\bar{x}\right\|^4.\end{aligned}$$

If $\|x\|^2 \geq \frac{626}{995}\|\bar{x}\|^2$,

$$\begin{aligned}\langle x, \mathbb{E}(\nabla f(x))\rangle - \frac{1}{500}\|x\|^2\|\bar{x}\|^2 &= 3\|x\|^4 - 2(x^\top\bar{x})^2 - \frac{501}{500}\|\bar{x}\|^2\|x\|^2 - \widetilde{\epsilon}\|x\|^2\\ &\geq 3\|x\|^4 - \frac{1}{2}\|x\|^2\|\bar{x}\|^2 - \frac{501}{500}\|\bar{x}\|^2\|x\|^2 - \widetilde{\epsilon}\|x\|^2\\ &\geq \frac{1}{100}\|x\|^4 + \frac{299}{100}\|x\|^4 - \frac{751}{500}\|\bar{x}\|^2\|x\|^2 - \widetilde{\epsilon}\|x\|^2\\ &\geq \frac{1}{100}\|x\|^4 + \frac{299}{100}\|x\|^4 - \left(\frac{751}{500} + \frac{1}{9\sqrt{2}}\right)\left(\frac{995}{626}\right)\|x\|^4\\ &\geq \frac{1}{100}\|\bar{x}\|^4,\end{aligned}$$

where we have used the fact that $\widetilde{\epsilon} \geq 0$ combined with the practical upper bound on c_s (4.6.13). Since $\frac{298}{451} \geq \frac{626}{995}$, we conclude that $\mathcal{R}_a \subset \mathcal{R}_1 \cup \mathcal{R}_2^x$.

- The set $\mathcal{R}_b \overset{\text{def}}{=} \left\{x \in \mathbb{R}^n : |x^\top\bar{x}| \geq \frac{1}{2}\|x\|\|\bar{x}\| \,;\, \|x\| \leq \frac{57}{100}\|\bar{x}\|\right\}$ is covered by the set $\mathcal{R}_1$. Indeed for any $x \in \mathcal{R}_b$ we have,

$$\begin{aligned}\left\langle \bar{x}; \nabla^2 f(x)\bar{x}\right\rangle + \frac{1}{100}\|\bar{x}\|^2\|x\|^2 &= 6(x^\top\bar{x})^2 + \frac{301}{300}\|x\|^2\|\bar{x}\|^2 - 3\|\bar{x}\|^4 - \widetilde{\epsilon}\|\bar{x}\|^2\\ &\leq \frac{901}{100}\|x\|^2\|\bar{x}\|^2 - \frac{149}{50}\|\bar{x}\|^4 - \widetilde{\epsilon}\|\bar{x}\|^2 - \frac{1}{50}\|\bar{x}\|^4\\ &\leq \frac{901}{100}\left(\frac{57}{100}\right)^2\|\bar{x}\|^2 - \frac{149}{50}\|\bar{x}\|^4 - \widetilde{\epsilon}\|\bar{x}\|^2 - \frac{1}{50}\|\bar{x}\|^4\\ &\leq -\widetilde{\epsilon}\|\bar{x}\|^2 - \frac{1}{50}\|\bar{x}\|^4 \leq -\frac{1}{50}\|\bar{x}\|^4,\end{aligned}$$

since $\widetilde{\epsilon} \geq 0$.

- Let consider the set $\mathcal{R}_c \overset{\text{def}}{=} \left\{x \in \mathbb{R}^n : \frac{1}{2}\|x\|\|\bar{x}\| \leq |x^\top\bar{x}| \leq \frac{99}{100}\|x\|\|\bar{x}\|\right\}$, which is covered by $\mathcal{R}_2^x$ and $\mathcal{R}_2^{h'}$. For any $x \in \mathcal{R}_c$ such that $\|x\| \geq \sqrt{\frac{1996}{1973}}\|\bar{x}\|$ we have

$$\begin{aligned}\langle x; \mathbb{E}(\nabla f(x))\rangle - \frac{1}{500}\|\bar{x}\|^2\|x\|^2 + \frac{1}{100}\|x\|^4 &= 3\|x\|^4 - 2(x^\top\bar{x})^2 - \|\bar{x}\|^2\|x\|^2 - \widetilde{\epsilon}\|x\|^2 - \frac{1}{500}\|\bar{x}\|^2\|x\|^2\\ &\geq \frac{299}{100}\|x\|^4 + \frac{501}{500}\|\bar{x}\|^2\|x\|^2 - 2(x^\top\bar{x})^2 - 2(x^\top\bar{x})^2 - \widetilde{\epsilon}\|x\|^2\\ &\geq \frac{299}{100}\|x\|^4 - \left(2\left(\frac{99}{100}\right)^2 + \frac{501}{500} + \frac{1}{9\sqrt{2}}\right)\|\bar{x}\|^2\|x\|^2\\ &\geq \left(\frac{299}{100}\left(\frac{1996}{1973}\right)^2 - \left(2\left(\frac{99}{100}\right)^2 + \frac{501}{500} + \frac{1}{9\sqrt{2}}\right)\right)\|x\|^4\\ &\geq 0.\end{aligned}$$

Therefore, we have $\mathcal{R}_c \cap \left\{x \in \mathbb{R}^n : \|x\| \geq \sqrt{\frac{1996}{1973}}\|\bar{x}\|\right\} \subset \mathcal{R}_2^x$. To show the remaining inclusion, we use an $(\alpha, \beta)-$type argument. Let assume that $\|x\| = \alpha\|\bar{x}\|$, $|x^\top\bar{x}| = \beta = \|\bar{x}\|\|x\| = \alpha\beta\|\bar{x}\|$ and $\widetilde{\epsilon} = \varepsilon\|\bar{x}\|^2$ with $\alpha \in \left[\frac{11}{20}, \sqrt{\frac{1996}{1973}}\right]$, $\beta \in \left[\frac{1}{2}, \frac{99}{100}\right]$ and $\varepsilon \in \left[0, \frac{1}{9\sqrt{2}}\right]$. We have

$$\begin{aligned}\langle x - \bar{x}, \mathbb{E}(\nabla f(x))\rangle &= 3\|x\|^4 + 3(x^\top\bar{x})\left(\|\bar{x}\|^2 - \|x\|^2\right) - 2(x^\top\bar{x})^2 - \|\bar{x}\|^2\|x\|^2 + \widetilde{\epsilon}\left((x^\top\bar{x}) - \|x\|\right)\\ &= \|\bar{x}\|^4\,\alpha\left(3\alpha^3 + 3\beta(1-\alpha^2) - 2\alpha\beta^2 - \alpha + \varepsilon(\beta - \alpha)\right).\end{aligned}$$

Whence we have

$$\begin{aligned}\frac{1}{\|\bar{x}\|^4\,\alpha}\left(\langle x - \bar{x}, \mathbb{E}(\nabla f(x))\rangle - \frac{1}{250}\|\bar{x}\|^2\|x\|\|d_x\|\right) &= 3\alpha^3 + 3\beta(1-\alpha^2) - 2\alpha\beta^2\\ &\quad - \alpha + \varepsilon(\beta - \alpha) - \frac{1}{250}\sqrt{1 + \alpha^2 - 2\alpha\beta}.\end{aligned}$$

It is straightforward that in this domain $\frac{1}{250}\sqrt{1+\alpha^2-2\alpha\beta} \leq \frac{1}{250}\sqrt{\frac{3969-2\sqrt{984527}}{1973}} \leq \frac{41}{10000}$. Therefore we define the following function

$$p(\alpha,\beta,\varepsilon) \stackrel{\text{def}}{=} 3\alpha^3+3\beta(1-\alpha^2)-2\alpha\beta^2-\alpha+\varepsilon(\beta-\alpha)-\frac{41}{10000}$$

Then p has a unique minimizer arising at $\left(0.998237, \frac{99}{100}, \frac{1}{9\sqrt{2}}\right)$ with a value $\frac{87239}{2500000}$. We deduce that

$$\langle x-\bar{x}, \mathbb{E}\left(\nabla f(x)\right)\rangle - \frac{1}{250}\left\|\bar{x}\right\|^2\left\|x\right\|\left\|d_x\right\| \geq 0$$

Therefore $\mathcal{R}_c \subset \mathcal{R}_2^x \cup \mathcal{R}_2^{h'}$.

- We now cover $\mathcal{R}_d \stackrel{\text{def}}{=} \left\{x\in\mathbb{R}^n : \frac{99}{100}\left\|\bar{x}\right\|\left\|x\right\| \leq |x^\top\bar{x}| \leq \left\|x\right\|\left\|\bar{x}\right\|, \left\|x\right\| \geq \frac{11}{20}\left\|\bar{x}\right\|\right\}$ with $\mathcal{R}_2^x, \mathcal{R}_3$ and $\mathcal{R}_2^{h'}$. For any $x\in\mathcal{R}_d$, with $\left\|x\right\| \geq \sqrt{\frac{1031}{1000}}\left\|\bar{x}\right\|$, we have

$$\begin{aligned}
\langle x, \mathbb{E}\left(\nabla f(x)\right)\rangle - \frac{1}{500}\left\|x\right\|^2\left\|\bar{x}\right\|^2 - \frac{1}{100}\left\|\bar{x}\right\|^4 &= \frac{299}{100}\left\|x\right\|^4 - 2(x^\top\bar{x})^2 - \frac{501}{500}\left\|\bar{x}\right\|^2\left\|x\right\|^2 - \widetilde{\epsilon}\left\|x\right\|^2 \\
&\geq \frac{299}{100}\left\|x\right\|^4 - \frac{1501}{500}\left\|\bar{x}\right\|^2\left\|x\right\|^2 - \widetilde{\epsilon}\left\|x\right\|^2 \\
&\geq \frac{299}{100}\left\|x\right\|^4 - \left(\frac{1501}{500}+\frac{1}{9\sqrt{2}}\right)\left\|\bar{x}\right\|^2\left\|x\right\|^2 \\
&\geq \left(\frac{299}{100} - \left(\frac{1501}{500}+\frac{1}{9\sqrt{2}}\right)\frac{1000}{1031}\right)\left\|x\right\|^4 \geq 0.
\end{aligned}$$

hence we have $\mathcal{R}_d \cap \left\{x\in\mathbb{R}^n : \left\|x\right\| \geq \sqrt{\frac{1031}{1000}}\left\|\bar{x}\right\|\right\} \subset \mathcal{R}_2^x$. When $\frac{23}{25} \leq \left\|x\right\| \leq \sqrt{\frac{1031}{1000}}$ we have,

$$\left\|d_x\right\|^2 = \left\|\bar{x}\right\|^2 + \left\|x\right\|^2 - 2x^\top\bar{x} = \left\|\bar{x}\right\|^2\left(1+\alpha^2-2\alpha\beta\right),$$

where we have $\alpha \in \left[\frac{23}{25}, \sqrt{\frac{1031}{1000}}\right]$ and $\beta \in \left[\frac{99}{100}, 1\right]$. Therefore, we consider

$$p(\alpha,\beta,\lambda) = 1+\alpha^2-2\alpha\beta-\frac{(1-\lambda)^2}{3},$$

with $\lambda \in \left]\frac{1}{9\sqrt{2}}, \frac{3}{5}\right]$. The maximum value of p is taken at $(\alpha,\beta,\lambda) = \left(\frac{23}{25}, \frac{99}{100}, \frac{3}{5}\right)$ thus $p(\alpha,\beta,\lambda) \leq -\frac{107}{3750}$. We deduce that $\mathcal{R}_d \cap \left\{x\in\mathbb{R}^n, \frac{23}{25} \leq \left\|x\right\| \leq \sqrt{\frac{1031}{1000}}\right\} \subset \mathcal{R}_3$. When $\frac{11}{20}\left\|\bar{x}\right\|\left\|x\right\| \leq \frac{24}{25}\left\|\bar{x}\right\|$, we have

$$\begin{aligned}
\frac{1}{\left\|\bar{x}\right\|^4\alpha}\left(\langle x-\bar{x}, \mathbb{E}\left(\nabla f(x)\right)\rangle - \frac{1}{250}\left\|\bar{x}\right\|^2\left\|x\right\|\left\|d_x\right\|\right) &= 3\alpha^3+3\beta(1-\alpha^2)-2\alpha\beta^2 \\
&\quad -\alpha+\varepsilon(\beta-\alpha)-\frac{1}{250}\sqrt{1+\alpha^2-2\alpha\beta}
\end{aligned}$$

where $\alpha \in \left[\frac{11}{20}, \sqrt{\frac{24}{25}}\right]$, $\beta \in \left[\frac{99}{100}, 1\right]$ and $\varepsilon \in \left[0, \frac{1}{9\sqrt{2}}\right]$. One check easily that

$$p(\alpha,\beta,\varepsilon) = 3\alpha^3+3\beta(1-\alpha^2)-2\alpha\beta^2-\alpha+\varepsilon(\beta-\alpha)-\frac{1}{250}\sqrt{1+\alpha^2-2\alpha\beta} \geq 0.$$

Consequently, we have $\mathcal{R}_d \cap \left\{x\in\mathbb{R}^n : \frac{11}{20}\left\|\bar{x}\right\|\left\|x\right\| \leq \frac{24}{25}\left\|\bar{x}\right\|\right\} \subset \mathcal{R}_2^{h'}$. Finally $\mathcal{R}_d \subset \mathcal{R}_2^x \cup \mathcal{R}_3 \cup \mathcal{R}_2^{h'}$.

By construction, we have $\mathcal{R}_a \cup \mathcal{R}_b \cup \mathcal{R}_c \cup \mathcal{R}_d = \mathbb{R}^n$, and therefore

$$\begin{aligned}
\mathbb{R}^n &= \mathcal{R}_a \cup \mathcal{R}_b \cup \mathcal{R}_c \cup \mathcal{R}_d \\
&\subset \mathcal{R}_1 \cup \mathcal{R}_2^x \cup \mathcal{R}_2^{h'} \cup \mathcal{R}_3 \\
&= \mathcal{R}_1 \cup \mathcal{R}_2^x \cup \left(\mathcal{R}_2^{h'} \cap \left\{ x \in \mathbb{R}^n : \frac{11}{20} \|\bar{x}\| \leq \|x\| \right\} \right) \cup \mathcal{R}_3 \\
&= \mathcal{R}_1 \cup \mathcal{R}_2^x \cup \left(\mathcal{R}_2^{h'} \cap \left\{ x \in \mathbb{R}^n : \frac{11}{20} \|\bar{x}\| \leq \|x\| \right\} \cap \mathcal{R}_3^c \right) \cup \mathcal{R}_3 \\
&= \mathcal{R}_1 \cup \mathcal{R}_2^x \cup \mathcal{R}_2^h \cup \mathcal{R}_3.
\end{aligned}$$

□

Part II

Phase Retrieval with Regularization

Chapter 5

Inertial Bregman Proximal Gradient

In this chapter, we study global and local convergence properties of an Inertial Bregman Proximal Gradient algorithm (IBPG) for minimizing the sum of two functions functions in finite dimension. One of the functions is assumed to be proper, closed, and convex but non-necessarily smooth whilst the second is a sufficiently smooth function but not necessarily convex. For the latter, we demand the smooth adaptable property *w.r.t* to some kernel/entropy which allows to remove the very popular global Lipschitz continuity requirement on its gradient. We consider IBPG under the framework of the triangle scaling property (TSP) which is a geometrical property for which one can provably ensure acceleration for a certain class of kernel/entropy functions in the convex setting. We provide global convergence guarantees when the kernel/entropy is strongly convex under the framework of the Kurdyka-Łojasiewicz property. Turning to the local convergence properties, we show that when the nonsmooth part is partly smooth relative to a smooth submanifold, IBPG has a finite activity identification property before entering a local linear convergence regime for which we establish a sharp estimate of the convergence rate. We report numerical simulations to illustrate our theoretical results on low complexity regularized phase retrieval.

In summary, the contributions of this chapter are:

Main contributions of this chapter

- Global convergence of IBPG under the Under the Kurdyka-Łojasiewicz property.
- Finite activity identification under partial smoothness.
- Local linear convergence analysis of IBPG with a sharp rate estimate.
- Saddle point escape property in the smooth case.

Contents

5.1 Introduction

5.1.1 Problem statement

In this work, we study the following class of composite non-necessarily smooth nor convex optimization problem

$$\inf_{x\in\mathbb{R}^n} \left\{ \Phi(x) \stackrel{\text{def}}{=} F(x) + G(x) \right\}, \tag{$\mathcal{P}$}$$

under the following assumptions:

Assumption 5.1.1.

(A.1) $F : \mathbb{R}^n \to \mathbb{R}$ *is* C^1-*smooth.*

(A.2) $G : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is a proper, lower semi-continuous and convex.*

(A.3) Φ *is bounded from below, i.e.* $\inf \Phi(\mathbb{R}^n) > -\infty$.

In many application in machine learning, data processing and inverse problems, the function G plays the role of a penalty/regularization term. It is intended to encode structural properties or prior knowledge information about the set of desired solutions. The function F on the other hand correspond to the loss function or the data fidelity term.

Throughout this chapter, the smooth function F can be nonconvex which allows us to cover a large class of problems arising in applications such as statistics, machine learning, and in particular quadratic inverse problems, *i.e.*, phase retrieval.

To solve ($\mathcal{P}$), we associate to the objective Φ a kernel or entropy function ψ *w.r.t* which F is relatively smooth. In this work, we propose Algorithm 5 which we coin Inertial Bregman Proximal Gradient (IBPG).

Algorithm 5: Inertial Bregman Proximal Gradient

Parameters: $\kappa \in]1, 2]$;
Initialization: $z_{-1} = x_{-1}, z_0 = x_0 \in \mathbb{R}^n$, $a_{-1} = a_0 = 1$, and $0 < \underline{a} < \overline{a} \leq 1$;
for $k = 0, 1, \ldots$ **do**

$$
\begin{aligned}
&y_k = z_k + a_k(x_k - z_k);\\
&x_{k+1} = (\nabla\psi + \gamma_k \partial G)^{-1}\left(\nabla\psi(y_k) - \gamma_k \nabla F(y_k)\right), \quad \gamma_k = \frac{a_k^{\kappa-1}}{L};\\
&z_{k+1} = x_k + a_k(x_{k+1} - x_k);\\
&\text{Choose } a_{k+1} \in [\underline{a}, \overline{a}] \quad \text{s.t.} \quad (a_{k+1}^{1-\kappa} + 1)^{1/\kappa}(1 - a_{k+1}) < a_k^{1/\kappa - 1}/(1 - a_k).
\end{aligned}
\tag{IBPG}
$$

I can be easily shown that y_k can be written as a recursion of (x_k, x_{k-1}) with inertial parameters that depend on (a_k, a_{k-1}). When $a_k \equiv 1$, we recover the Bregman Proximal Gradient (BPG without inertia) whose global convergence was already established in [40].

In [89], a slightly different algorithm called Accelerated Bregman Proximal Gradient, was considered in the convex case. This scheme was inspired by the Improved Interior Gradient Algorithm for conic optimization as developed in [13, Section 5]. It was shown in [89] that the Accelerated Bregman Proximal Gradient indeed provides acceleration when the kernel ψ satisfies the triangle scaling property (TSP) (see Definition 2.3.9) with a convergence rate on the values of $O(k^{-\kappa})$, where $\kappa \in]1, 2]$ is the triangle scaling exponent (TSE). For $\kappa = 2$, one recovers the standard Nesterov-like accelerated rate.

In the nonconvex case, another inertial scheme was analyzed in [136] when the Bregman kernel/entropy is also strongly convex and G is weakly (or semi-) convex. Global convergence of the iterates under KL was proved there using line search on both the extrapolation (inertial) parameter and the descent step-size. Using more structure of the kernel, *i.e.* the TSP, will allow to have a sharper analysis and results.

5.1.2 Contributions

In this chapter, we study the global and local convergence properties of Algorithm 5 for a class of Bregman kernels that satisfies the TSP property. The main contributions of this work are:

Global convergence of the scheme Under the Kurdyka-Łojasiewicz (KL) property, we show that bounded iterates generated by the Inertial Bregman Proximal Gradient converge to a critical point of the objective function when the inertial parameters satisfy the condition that $a_k \in [\overline{a}, \underline{a}] \subset]0, 1[$. We also show that starting near an optimal solution, the sequence converges to it.

Finite activity identification We establish that IBPG enjoys a finite activity identification property. More precisely, we show that when the nonsmooth part G is partly smooth with respect to an underlying manifold $\mathcal{M}_{x_\star}$ near some critical point $x_\star$, the iterates will identify the manifold under an appropriate nondegeneracy condition. This identification phenomenon implies the existence of some large number K such that the iterates generated after this number lie in the manifold $\mathcal{M}_{x_\star}$.

The nondegeneracy condition cannot be relaxed in general as shown in [92, Example 4.1, 4.2, 4.3] respectively for the projected gradient descent, Newton method, and the proximal point algorithm.

Local linear convergence analysis After activity identification property, we show that locally along the active manifold $\mathcal{M}_{x_\star}$, we can linearize the iterates generated by IBPG when F is locally C^2. In the case, we provide a spectral analysis of the linearized system, and under appropriate restricted injectivity, we exhibit a linear convergence regime for proper choice of the inertial parameter. This choice depends in particular on the TSE parameter κ, which generalizes the Euclidean case for which this parameter is just 2. Our work hence extends that of [117] to non-euclidean Bregman-based geometry.

Escape property in the smooth case Equipped with the center stable manifold theorem, we study the trap avoidance of the inertial Bregman gradient method where $G \equiv 0$. We show that the scheme generically avoids strict saddle points, which are critical points where the function has at least one direction of negative curvature.

5.2 Global Convergence Analysis

5.2.1 Main assumptions

We will need the following assumptions on ψ, which will be invoked jointly or separately in our proofs.

Assumption 5.2.1.

(B.1) *ψ is a C^2 σ_ψ-strongly convex function with $\sigma_\psi > 0$.*

(B.2) *F is $L-$smooth relative to ψ on $\mathbb{R}^n$.*

(B.3) *∇F and $\nabla\psi$ are Lipschitz continuous on bounded subsets of $\mathbb{R}^n$.*

Remark 5.2.2.

- Strong convexity of ψ plays an important role to establish global convergence of the sequences of iterates.
- C^2 smoothness of ψ is only needed occasionally and some of our statements remain true even without it.
- Assumption (B.1) implies that ψ is Legendre and thus $\nabla\psi$ is a bijection on $\mathbb{R}^n$ whose inverse is $\nabla\psi^*$ (see Remark 2.3.2). Moreover, strong convexity implies that $\nabla^2\psi^*(\nabla\psi(x)) = \nabla^2\psi(x)^{-1}$; see *e.g.*, [106, Lemma 2.2].
- Assumption (A.2) together with (B.1) imply that the D-prox operator of index $\gamma > 0$

$$(\nabla\psi + \gamma\partial G)^{-1} \circ \nabla\psi : x \in \mathbb{R}^n \mapsto \operatorname*{Argmin}_{z\in\mathbb{R}^n} G(z) + \frac{1}{\gamma} D_\psi(z, x)$$

  is single-valued; see [26, Proposition 3.22]. Strong convexity of ψ can be weakened to strict convexity and legenderness if $\psi + \gamma G$ is supercoercive. If convexity of G is removed, the D-prox is nonempty and compact-valued if ψ is Legendre and $\psi + \gamma G$ is supercoercive; see [40, Lemma 3.1].
- Our analysis and results can be extended to handle the constrained case where ($\mathcal{P}$) is solved over a closed convex set. This necessitates that ψ to be a barrier function of the constraint set and some technical (domain) adaptations of our assumptions that we prefer to avoid here for the sake of clarity[1].

[1] Anyway, our focus being on phase retrieval, our current setting is sufficient.

5.2.2 Convergence analysis

Our main result states that if Φ is also a KL function, then bounded iterates of IBPG converge to a critical point of Φ.

Theorem 5.2.3. *Consider problem ($\mathcal{P}$) under Assumptions 5.1.1-5.2.1. Suppose that the sequence of IBPG parameters $(a_k)_{k\in\mathbb{N}}$ are chosen as in Algorithm 5. Then,*

(i) $\sum_{k\in\mathbb{N}} \|x_k - x_{k-1}\|^2 < \infty$ *and*

$$\min_{0\leq i\leq k} \|x_i - x_{i-1}\|^2 \leq \frac{2(\sigma_\psi \nu L)^{-1}\Psi_0(x_0, x_{-1})}{k+1}.$$

Assume moreover that Φ and ψ satisfy the KL property, that $a_k \equiv a \in [\underline{a}, \overline{a}]$ for all $k \geq K$, where K is arbitrarily large. If the sequence of IBPG iterates $(x_k)_{k\in\mathbb{N}}$ is bounded then,

(ii) all sequences $(x_k)_{k\in\mathbb{N}}$, $(y_k)_{k\in\mathbb{N}}$ and $(z_k)_{k\in\mathbb{N}}$ have finite length and converge to the same limit in $\mathrm{crit}(\Phi)$.

(iii) If $\mathrm{Argmin}(\Phi) \neq \emptyset$ *and IBPG is started near a global minimizer $x^\star$ in the $\Phi-$attentive topology, then the generated sequences converge to $x^\star$.*

We defer the proof to Section 5.6.

Remark 5.2.4.

- The boundedness of the sequence is a standard assumption for the global convergence of the sequence in the nonconvex case. Coercivity of Φ is for instance sufficient to ensure it.
- Choosing a_k constant for k large enough is a standard strategy for the Lyapunov analysis in the nonconvex case. A similar strategy is also used in [39] and [136]. This also makes sense in practice, a fixed inertial parameter as in the heavy ball method with friction is popular for inertial algorithms.

The choice of $(\gamma_k)_{k\in\mathbb{N}}$ and $(a_k)_{k\in\mathbb{N}}$ devised in Algorithm 5 is sufficient for our analysis. Actually, we only need that

$$\gamma_k \in]0, 1/L] \text{ and } (L + \gamma_k^{-1})(1-a_k)^\kappa(1-a_{k-1})^\kappa < \gamma_{k-1}^{-1}.$$

For instance, if $\gamma_k \equiv 1/L$, then it is sufficient that $(1-a_k) < 2^{-1/\kappa}/(1-a_{k-1})$ which is easy to verify since $a_k \in]0, 1]$.

There are many possible choices of the sequence $(a_k)_{k\in\mathbb{N}}$ that obey the condition of Algorithm 5. For instance, take $a_k = \frac{k+1}{k+1+\alpha} \in [1/(1+\alpha), 1]$, where $\alpha > 0$. To verify that the condition holds, observe that $(a_{k+1}^{1-\kappa} + 1)(1-a_{k+1})^\kappa$ is decreasing in k. On the other hand, the function $h : [1/(1+\alpha), 1] \mapsto a^{1-\kappa}/(1-a)^\kappa$ has a unique minimum on $[0, 1]$ at $a_{\min} = \max(1/(1+\alpha), (\kappa-1)/(2\kappa-1))$. Thus, for the inequality to hold true, it is sufficient that $(a_1^{1-\kappa} + 1)(1-a_1)^\kappa < h(a_{\min})$ for all $\kappa \in]1, 2]$. This is achieved by taking $\alpha \leq 3$.

5.3 Local Convergence Analysis

In this section, we present the local analysis of the Inertial Bregman Proximal Gradient. We start with the following definition.

Definition 5.3.1. (Nondegenerate critical point for composite function) We say that a critical point satisfies the nondegeneracy condition for the composite function Φ if:

$$-\nabla F(x_\star) \in \mathrm{ri}(\partial G(x_\star)). \tag{ND}$$

Let $\mathrm{ND}(\Phi)$ denote the set of critical points satisfying this condition for Φ.

Remark 5.3.2.

(i) Applying [75, Proposition 10.12], it turns out that for $G \in \Gamma_0(\mathbb{R}^n)$, the notion of identifiable manifold is equivalent to partial smoothness around an active smooth manifold combined with the nondegeneracy condition.

(ii) Suppose that Φ is a semi-algebraic function, and more generally a function definable on an o-minimal structure. It follows from [74, Theorem 4.16] that generically on $v \in \mathbb{R}^n$, the function $\Phi_v \overset{\text{def}}{=} \Phi(x) - \langle v, x \rangle$ has a finite number of critical points and each critical point is nondegenerate and admits an identifiable manifold. In plain words, assumption (ND) is generic.

5.3.1 Finite activity identification of IBPG

The following result shows that IBPG generates a sequence that identifies active manifolds in finite time.

Lemma 5.3.3 (Finite time activity identification). *Let us consider an instance of Algorithm 5 such that $(x_k)_{k\in\mathbb{N}}$ is bounded. Let $x_\star \in \operatorname{crit}(\Phi)$ be the limit of the sequence and assume that $G \in \mathrm{PSF}_{x_\star}(\mathcal{M}_{x_\star})$ with $x_\star \in \mathrm{ND}(\Phi)$. Under the same assumptions as Theorem 5.2.3, there exists a constant K large enough such that for all $k \geq K, x_k \in \mathcal{M}_{x_\star}$.*

(i) $\mathcal{M}_{x_\star}$ is an affine subspace, then $\mathcal{M}_{x_\star} = x_\star + T_{x_\star}$ and $(y_k)_{k\in\mathbb{N}}, (z_k)_{k\in\mathbb{N}} \in \mathcal{M}_{x_\star}, k \in K$,

(ii) If moreover, G is locally polyhedral around $x_\star$ then for all $k \geq K$, the remaining sequences satisfy $y_k, z_k \in \mathcal{M}_{x_\star}$ and $\nabla^2_{\mathcal{M}_{x_\star}} G(x_k) = 0$.

See the Section 5.7.1 for the proof of this lemma.

5.3.2 Local linearization of IBPG

In this section, let us consider $x_\star$ a critical point of Φ and let $\mathcal{M}_{x_\star}$ be a C^2-smooth submanifold such that $G \in \mathrm{PSF}_{x_\star}(\mathcal{M}_{x_\star})$. Let us assume that the smooth part F is C^2 around $x_\star$. Let us denote $T_{x_\star} \overset{\text{def}}{=} \mathcal{T}_{\mathcal{M}}(x_\star)$ and fix a stepsize $\gamma_k \in]0, 1/L]$. For the rest of the analysis, we define the following matrices which help us to capture the local behavior of the iterates.

$$H_F \overset{\text{def}}{=} \gamma \mathrm{P}_{T_{x_\star}} \nabla^2 F(x_\star) \mathrm{P}_{T_{x_\star}}, \quad H_\psi \overset{\text{def}}{=} \mathrm{P}_{T_{x_\star}} \nabla^2 \psi(x_\star) \mathrm{P}_{T_{x_\star}}, \quad V \overset{\text{def}}{=} H_\psi - H_F, \tag{5.3.1}$$
$$U \overset{\text{def}}{=} \gamma \nabla^2_{\mathcal{M}_{x_\star}} \Phi(x_\star) \mathrm{P}_{T_{x_\star}} - H_F.$$

where $\nabla^2_{\mathcal{M}_{x_\star}} \Phi$ denotes the Riemannian Hessian of Φ along the submanifold $\mathcal{M}_{x_\star}$ and $\mathrm{P}_{T_{x_\star}}$ the projection onto $T_{x_\star}$.

Remark 5.3.4. Since $G \in \Gamma_0(\mathbb{R}^n)$, [118, Lemma 4.3] shows that U is symmetric positive semi-definite under the condition that $x_\star \in \mathrm{ND}(\Phi)$ or that $\mathcal{M}_{x_\star}$ is an affine space. Therefore $H_\psi + U$ is symmetric positive definite and hence invertible. Let us denote

$$W \overset{\text{def}}{=} (H_\psi + U)^{-1}.$$

We have that W is symmetric positive with eigenvalues in $]0, 1/\sigma_\psi]$.

Definition 5.3.5. (Restricted injectivity) We say that a critical point $x_\star$ satisfies the restricted injectivity condition if there exists $\sigma \geq 0$ such that the following condition holds true

$$\forall h \in T_{x_\star}, \quad \left\langle h, \left(\nabla^2 F(x_\star) - \sigma \nabla^2 \psi(x_\star)\right) h \right\rangle \geq 0. \tag{5.3.2}$$

We denote by $\mathrm{RI}(\Phi)$ the set of all critical points where the restricted injectivity is satisfied. In this case, the local continuity of the Hessian of F implies that $\ker(\nabla^2 F(x_\star)) \cap T_{x_\star} = \{0\}$. Observe that

by [74, Theorem 4.16] and strong convexity of ψ, (5.3.2) is equivalent to the fact that $x_\star$ is a stable strong local minimizer of Φ.

Let $a \in [\underline{a}, \overline{a}]$ and define $r_k \overset{\text{def}}{=} x_k - x^\star$ and $d_k \overset{\text{def}}{=} \begin{pmatrix} r_k \\ r_{k-1} \end{pmatrix}$, we will need the following key matrix

$$M \overset{\text{def}}{=} \begin{bmatrix} (2a - a^2)WV & (1-a)^2 WV \\ \mathrm{Id} & 0 \end{bmatrix}. \tag{5.3.3}$$

At this step, we can describe the local behavior of the sequence generated by Algorithm 5. The next result is a local linearization of the iterative scheme.

Proposition 5.3.6. *Consider the problem $(\mathcal{P})$ under Assumptions 5.1.1-5.2.1. Let us assume that the sequences produced by Algorithm 5 converge to $x_\star \in \mathrm{ND}(\Psi) \cap \mathrm{RI}(\Psi)$ with $G \in \mathrm{PSF}_{x_\star}(\mathcal{M}_{x_\star})$. If F is C^2 locally around $x_\star$ and the inertial parameter sequence $(a_k)_{k\in\mathbb{N}}$ satisfies $a_k \to a$ then for k large enough, we have*

$$d_{k+1} = M d_k + o(\|d_k\|). \tag{5.3.4}$$

The little "o" term disappears when G is locally polyhedral and a_k is chosen constant.

See Section 5.7.2 for the proof of this proposition.

Remark 5.3.7.

(i) If $a_k \equiv 1$, we recover the Bregman Proximal Gradient and we have the following linearized iteration

$$r_{k+1} = WV r_k + o(\|r_k\|).$$

(ii) When the kernel is the energy, *i.e.* $\psi = \|\cdot\|^2/2$, a similar analysis has been done for a symmetric version of the inertial Forward-Backward [118, Proposition 4.5] with a different choice of inertial parameters. Our result is however different and it involves the new matrix H_ψ which makes the spectral analysis of M more intricate.

5.3.3 Spectral properties of M

Our goal now is to show local linear convergence of IBPG. Towards this, we examine the structure of the locally linearized iteration given in (5.3.4). It is adequate to upper-bound (strictly) the spectral radius of M by 1 and subsequently draw conclusions using standard reasoning. We will relate the eigenvalues of M to those of WV. Let η and ϱ be an eigenvalue of WV and M respectively. We denote $\underline{\eta}$ and $\overline{\eta}$ as the smallest and largest (signed) eigenvalues of WV, and $\rho(M)$ as the spectral radius of M. When G is a general partly smooth function, then U is nontrivial, we have the following proposition.

Proposition 5.3.8. *Let us define $\Lambda \overset{\text{def}}{=} |q_\psi(x_\star) - \gamma\sigma|$ where $q_\psi(x_\star) = \frac{\lambda_{\max}(\nabla^2\psi(x_\star))}{\sigma_\psi}$. Denote $q_F(x_\star) = \frac{L}{\sigma}$.*

(i) Let $\begin{pmatrix} r_1 \\ r_2 \end{pmatrix}$ be an eigenvector of M corresponding to an eigenvalue ϱ then it must satisfy $r_1 = \varrho r_2$. Besides, r_2 is an eigenvector of WV associated with the eigenvalue η, where η and ϱ satisfy the relation

$$\varrho^2 - (2a - a^2)\varrho\eta - (1-a)^2\eta = 0, \tag{5.3.5}$$

and $\rho(M) \leq \rho(WV) < \Lambda$ if, and only if,

$$\frac{1}{2a^2 - 4a + 1} < \underline{\eta}.$$

(ii) If the inertial parameters are chosen such that $a \in [\underline{a}, \overline{a}]$ *with* $\underline{a} > \sqrt[\kappa-1]{q_F(x_\star)(q_\psi(x_\star)-1)}$ *and* $\overline{a} < \sqrt[\kappa-1]{q_F(x_\star)(q_\psi(x_\star)+1)}$ *then* $\Lambda < 1$.

See section 5.7.3 for the proof.

Remark 5.3.9.

(ii) In the euclidean case, $\Lambda = \left(1 - a_k^{\kappa-1}\frac{\sigma}{L}\right) < 1$ since $\sigma \leq L$ and $a_k \in]0,1]$.

(iii) Claim (ii) states that $\rho(M) < 1$ whenever a is small enough while being bounded away from zero.

5.3.4 Local linear convergence

We can now state the local linear convergence result.

Theorem 5.3.10. ***(Local linear convergence)*** *Consider the problem* $(\mathcal{P})$ *under Assumptions 5.1.1-5.2.1. Let* $(x_k)_{k\in\mathbb{N}}$ *be a bounded sequence produced by Algorithm 5 that converges to* $x_\star \in \mathrm{ND}(\Phi)\cap\mathrm{RI}(\Phi)$ *with* $G \in \mathrm{PSF}_{x_\star}(\mathcal{M}_{x_\star})$*, and assume that* F *is* C^2 *locally around* $x_\star$*. If* a *is such that Proposition 5.3.8(ii) holds, then* $(x_k)_{k\in\mathbb{N}}$ *converges locally linearly to* $x_\star$*. More precisely, given any* $\rho \in [\rho(M), 1[$*, there exist* $K \in \mathbb{N}$ *large enough such that* $\forall k \geq K$ *,*

$$\frac{\|z_k - x_\star\|}{\|z_K - x_\star\|} = O(\rho^{k-K}). \tag{5.3.6}$$

Proof. First use the global convergence result combined with the local linearization Proposition 5.3.6 and the spectral analysis of M in Proposition 5.3.8. Then conclude by standard arguments. □

5.4 Escape Property in the Smooth Case

5.4.1 Trap avoidance for the inertial mirror descent

Throughout this subsection, we assume that $G \equiv 0$. Thus IBPG reduces to the Inertial Mirror Descent (IMD) which is a variant of the Improved Interior Gradient Algorithm [13]. We assume that the algorithm is run with a fixed stepsize and a fix inertial parameter. The scheme is summarized in Algorithm 6 for the reader's convenience.

Algorithm 6: Inertial Mirror Descent

Parameters: $\kappa \in]1,2]$;
Initialization: $z_{-1} = x_{-1}, z_0 = x_0 \in \mathbb{R}^n$, $a_{-1} = a_0 = 1$, and fix $a \in]0,1]$;
for $k = 0, 1, \ldots$ **do**

$$y_k = z_k + a(x_k - z_k);$$
$$x_{k+1} = \nabla\psi^{-1}\left(\nabla\psi(y_k) - \gamma\nabla F(y_k)\right), \quad \gamma = \frac{a^{\kappa-1}}{L};$$
$$z_{k+1} = x_k + a(x_{k+1} - x_k).$$

Theorem 5.4.1. ***(Trap avoidance of IMD.)*** *Consider the minimization problem* $(\mathcal{P})$ *with* $G \equiv 0$ *under Assumption 5.1.1-5.2.1 and let* $x_\star \in \mathrm{crit}(\Phi)$*. Then for almost all initilizers* (x_0, x_{-1}) *of IMD, the generated sequences converge to a critical point that is not a strict saddle point.*

We defer the proof to Section 5.8.1. Clearly, this means that if (x_0, x_{-1}) is drawn at random from a distribution with has a density *w.r.t* Lebesgue measure, then with probability one, IMD converges to a critical point which is not a strict saddle.

5.4.2 Challenges of the escape property for IBPG in the nonsmooth case

In this section, we discuss the difficulties and challenges posed by the case where $G \neq 0$ in $\mathcal{P}$ is nonsmooth. First one has to adapt the notion of strict saddles to the nonsmooth setting. Adopting the terminology in [68], we introduce the following notion of active strict saddles.

Definition 5.4.2 (Active strict saddle). Let us consider $g : \mathbb{R}^n \to \overline{\mathbb{R}}$. We say that a point $x_\star$ is an active strict saddle point of the nonsmooth function g if

(i) $x_\star \in \mathrm{crit}(g)$ *i.e.*, $0 \in \partial g(x_\star)$.

(ii) There exists an active manifold $\mathcal{M}$ at the point $x_\star$.

(iii) The Riemannian Hessian of g at $x_\star$ has a at least one negative eigenvalue.

Let us denote by $\mathrm{Actstrisad}(g)$ the set of all active strict saddle points of g.

In the euclidean setting, the authors in [68] showed that proximal methods with weakly convex and definable functions generically avoid active strict saddle points. At the core of their proofs is again the center stable manifold theorem. In turn, the regularity required by this theorem heavily rely on the properties of the proximal mapping in the euclidean setting, and in particular its firm nonexpansiveness, as well as Lipschitz continuity of the gradient of the smooth part.

In the Bregman setting, these properties are not true anymore. In fact, the proof strategy consists in characterizing the regularity and the spectrum of the the following fixed point mapping that characterizes Algorithm 5,

$$\mathbf{T}(x_2, x_1) = \begin{bmatrix} (\nabla\psi + \gamma\partial G)^* \left(\nabla\psi(y(x_2, x_1)) - \gamma\nabla F(y(x_2, x_1))\right) \\ x_2 \end{bmatrix},$$

where

$$a \in [\underline{a}, \overline{a}] \text{ and } y(x_2, x_1) = (2a - a^2)x_2 + (1-a)^2 x_1.$$

One has that $x_\star \in \mathrm{crit}(\Phi)$ if and only if $(x_\star, x_\star)$ is a fixed point of the operator $\mathbf{T}$. We have the following result.

Lemma 5.4.3. *Consider the minimization problem ($\mathcal{P}$) under Assumptions 5.1.1-5.2.1 and let $x_\star \in \mathrm{crit}(\Phi)$. Then $\mathbf{T}$ is a C^1-smooth function in a neighborhood of $(x_\star, x_\star)$. Besides, if $x_\star$ is an active strict saddle of Φ then the jacobian $D\mathbf{T}(x_\star, x_\star)$ has a real eigenvalue that is strictly greater than one.*

We defer the proof to Section 5.8.2.

This lemma extends [68, Theorem 4.1] to the inertial Bregman proximal gradient setting. The key insight is that, in the vicinity of the critical point denoted by $x_\star$, any optimization problem can be locally approximated on the active manifold. This allows us to circumvent the nonsmoothness issues present in the general case.

However, extending [68, Theorem 4.1] to the Bregman setting presents a significant challenge. In the Euclidean framework, the arguments rely heavily on [68, Corollary 2.12] of the center manifold theorem. This result hinges on the assumption that the mapping A is a global lipeomorphism. Unfortunately, this is not true in the Bregman setting, as A is only a local lipeomorphism but not a global one. This suggests that a different proof strategy is needed which calls for future work as we will discuss in our perspectives.

5.5 Numerical Experiments

In this section, we discuss some numerical experiments to illustrate our theoretical results.

5.5.1 Phase retrieval

We apply our results to regularized phase retrieval. Recall that the goal is to recover a vector $\bar{x} \in \mathbb{R}^n$ from quadratic measurements

$$y = |A\bar{x}|^2 \in \mathbb{R}^m,$$

where $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator. We can reformulate this problem as an optimization problem in the form

$$\min_{x\in\mathbb{R}^n} \Phi(x) = \frac{1}{4m} \left\| y - |Ax|^2 \right\|^2 + \lambda R(x), \quad \lambda > 0. \tag{5.5.1}$$

where $R \in \Gamma_0(\mathbb{R}^n)$. R is a regularizer that that promotes objects sharing a structure similar to that of $\bar{x}$. Problem (5.5.1) is an instance of $(\mathcal{P})$ with $F(x) = \frac{1}{4m} \left\| y - |Ax|^2 \right\|^2$ and $G(x) = \lambda R(x)$.

Let us observe that F is a semi-algebraic data fidelity term and that $F \in C^2(\mathbb{R}^n)$ but is nonconvex (though weakly convex). Besides, ∇F is not Lipschitz continuous. Therefore, we associate to F the kernel function

$$\psi(x) = \frac{1}{4} \|x\|^4 + \frac{1}{2} \|x\|^2 . \tag{5.5.2}$$

$\psi \in C^2(\mathbb{R}^n)$ is full domain and 1−strongly convex function with a gradient that is Lipschitz over bounded subsets of $\mathbb{R}^n$; see Proposition 3.2.2. We have already seen that F is smooth relative to ψ (see Lemma 3.2.3). Thus all our Assumptions 5.1.1-5.2.1 are fulfilled.

5.5.2 Experiments setup

Throughout our experiments, A is drawn from the standard Gaussian ensemble, *i.e.*, the entries of A are i.i.d mean-zero and standard Gaussian. We solve this problem using Algorithm 5 with $a_k \equiv 1$. For each numerical experiment, we run the algorithm with a constant step-size $\gamma = \frac{0.99}{3+10^{-4}}$. For R, we have tested several regularizers as described hereafter. All the results on partial smoothness of this part are taken from [174].

Example 5.5.1. (ℓ_1-norm). For any $x \in \mathbb{R}^n$, the ℓ_1-norm is given by $R(x) = \|x\|_1 \overset{\text{def}}{=} \sum_{i=1}^n |x_i|$, which is partly smooth at any x relative to the linear subspace

$$\mathcal{M} = T_x \overset{\text{def}}{=} \{u \in \mathbb{R}^n : \operatorname{supp}(u) \subset \operatorname{supp}(x)\}, \quad \operatorname{supp}(x) \overset{\text{def}}{=} \{i : x_i \neq 0\}.$$

The underlying vector $\bar{x}$ is taken to be sparse with $s = 12$ non-zeros entries for a vector of size $n = 128$. The number of quadratic measurements is taken as $m = 0.5 \times s^{1.5} \times \log(n)$, which is in line with the bounds discussed in Chapter 6. As there is no noise, we took $\lambda = 10^{-8}$. Figure 5.1 shows the recovery results. The left plot of Figure 5.1 displays the relative error of the iterates vs the number of iterations. On the right plot, we display the support of the iterates. Clearly, the left plot shows that Algorithm 5 identifies the correct support after 300 iterations and converges to the true vector. The left plot confirms what is anticipated by our analysis, that the relative error converges locally linearly (see the dashed line). The local linear convergence rate is in very good agreement with the one we predicted.

Example 5.5.2. ($\ell_{1,2}$-norm). Here, we take R as the group/block Lasso which si designed to promote group sparsity. Let $\{1, \cdots, n\}$ be partitioned into nonoverlapping blocks $\mathcal{B}$ such that: $\bigcup_{b\in\mathcal{B}} = \{1, \cdots, n\}$. The $\ell_{1,2}$-norm of x is given by $R(x) = \|x\|_{1,2} \overset{\text{def}}{=} \sum_{b\in\mathcal{B}} \|x_b\|$ with $x_b = (x_i)_{i\in b} \in \mathbb{R}^{|b|}$. This function is partly smooth at x with respect to the linear subspace

$$\mathcal{M} = T_x \overset{\text{def}}{=} \left\{u \in \mathbb{R}^n : \operatorname{supp}_{\mathcal{B}}(u) \subset \mathcal{S}_{\mathcal{B}}, \quad \mathcal{S}_{\mathcal{B}} \overset{\text{def}}{=} \bigcup \{b : x_b \neq 0\}\right\}.$$

In our experiment, we consider the true vector is of size $n = 128$ with 2 nonzero blocks of size 8 each. The number of measurements is $m = 0.5 \times (2 \times 8)^2 \times \log(128)$ quadratic measurements which is

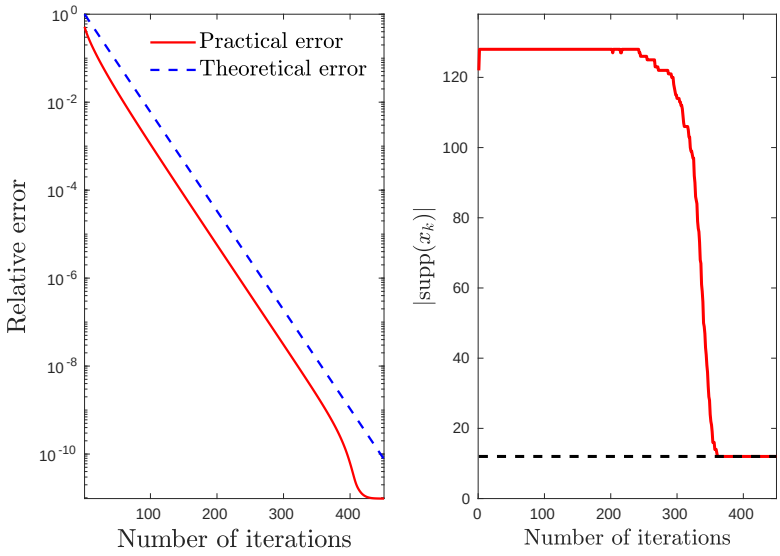


Figure 5.1: Phase retrieval by solving (5.5.1) with the ℓ_1-norm regularizer.

again in agreement with our bounds in Chapter 6. We also take $\lambda = 10^{-8}$. The results are shown in Figure 5.2, and they are consistent with the discussion for the ℓ_1-norm.

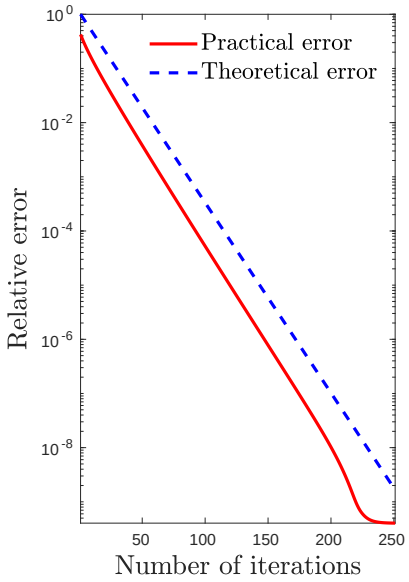


Figure 5.2: Phase retrieval by solving (5.5.1) with the $\ell_{1,2}-$norm regularizer.

Example 5.5.3. (Analysis-type prior). Let $R_0 \in \mathrm{PSF}_{Dx}(\mathcal{M}_0)$ where $D : \mathbb{R}^n \to \mathbb{R}^p$ is a linear operator. When D satisfies an appropriate transversality condition, then $R \stackrel{\text{def}}{=} R_0 \circ D$ is partly smooth with respect to $\mathcal{M} = \{u \in \mathbb{R}^n : Du \in \mathcal{M}_0\}$.

The anisotropic total variation is a particular case where R_0 is the ℓ_1-norm and D is a finite-difference operator with appropriate boundary conditions. It is is polyhedral and partly smooth at x relative to the linear subspace

$$\mathcal{M} = T_x \stackrel{\text{def}}{=} \{u \in \mathbb{R}^n : \mathrm{supp}(Du) \subset \mathrm{supp}(Dx)\}.$$

In our experiment here, the original vector $\bar{x}$ is piecewise constant with $s = 12$ randomly placed jumps. The number of measurements is $m = 0.5 \times s^2 \times \log(n)$. The regularizer is the total variation. Since the proximity operator of the latter is not explicit, we used the maxflow algorithm of [59] to compute it.

The results are depicted in Figure 5.3. The left plot shows the original (dashed line) and the recovered vector (solid line). The right plot shows the evolution of the relative error vs iterations where again, a linear convergence behaviour is observed with a predicted rate that is very close to the observed one.

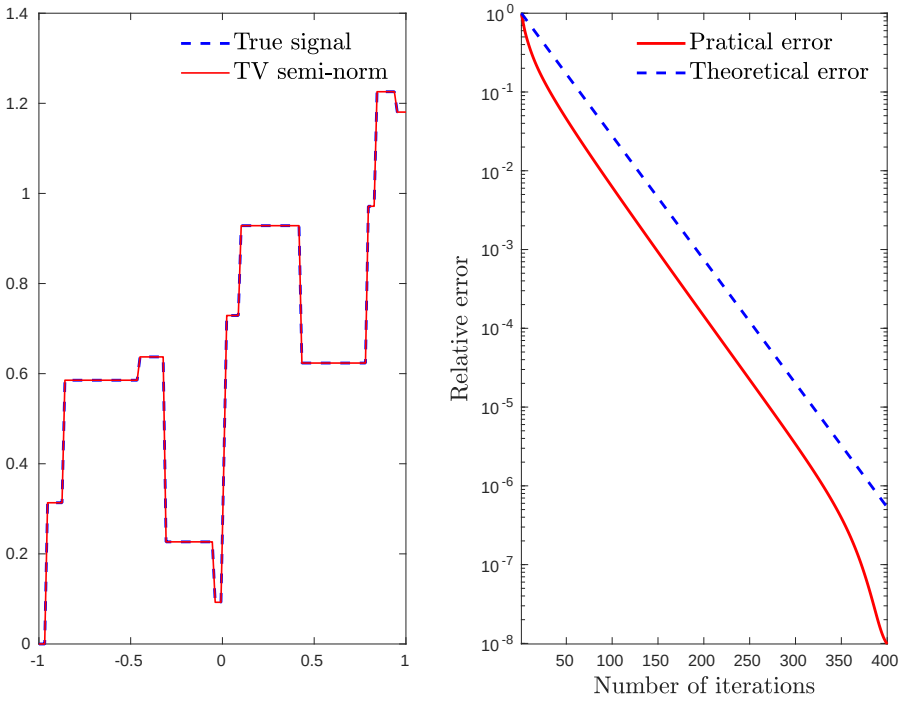


Figure 5.3: Phase retrieval by solving (5.5.1) with the TV semi-norm.

Example 5.5.4. (Wavelet synthesis-type prior). We here cast the phase retrieval problem as

$$\min_{v\in\mathbb{R}^p} \Phi(v) \stackrel{\text{def}}{=} \frac{1}{4m}\left\| y - |AWv|^2 \right\|^2 + \lambda \|v\|_1, \quad \lambda > 0, \tag{5.5.3}$$

where W is a wavelet synthesis operator. The reconstructed vector is given by $x = Wv$. When W is orthonormal, this is equivalent to the analysis-type formulation with $D = W^\top$. This is not anymore the case when W is redundant.

In this experiment, we will use the shift-invariant wavelet dictionary with the Haar wavelet, which is closely related to the total variation regularizer for 1D signals; see [167]. We take the same number of jumps and measurements as in the previous example. The results are shown in Figure 5.4.

5.6 Proof of Global Convergence

Lemma 5.6.1. *Let $(x_k)_{k\in\mathbb{N}}, (y_k)_{k\in\mathbb{N}}$ and $(z_k)_{k\in\mathbb{N}}$ be generated by IBPG (Algorithm 5). Then the following holds true*

(i) If $(x_k)_{k\in\mathbb{N}}$ is bounded, then the other sequences are also bounded.

(ii) If $(x_k)_{k\in\mathbb{N}}$ has a limit, then the remaining sequences also converge to the same limit.

Proof. (i) z_k being a convex combination of x_k and x_{k-1}, the conclusion is immediate. The same holds also for y_k.

(ii) Suppose that $x_k \to x^*$. We have by definition

$$\|z_k - x^*\| \leq \|x_k - x^*\| + \|x_k - x_{k-1}\|.$$

Passing to the limit we get that $z_k \to x^*$. Moreover,

$$\|y_k - x^*\| \leq (1 - a_k)\|z_k - x^*\| + a_k\|x_k - x^*\| \leq \|z_k - x^*\| + \|x_k - x^*\|.$$

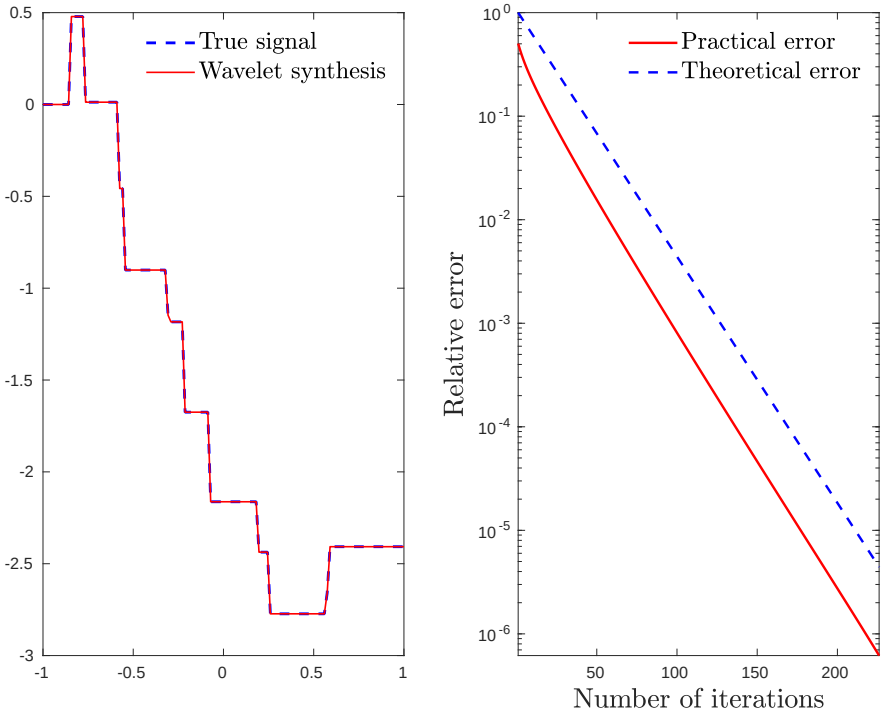


Figure 5.4: Phase retrieval with the synthesis prior formulation.

and thus $y_k \to x^*$.

□

Our global convergence analysis will be based on a Lyapunov analysis with the energy function Ψ_k on $\mathbb{R}^{3n}$ defined as

$$\Psi_k(x_k, x_{k-1}) = \Phi(x_k) - \inf \Phi + a_{k-1}^{1-\kappa} L D_\psi(x_k, x_{k-1}),$$

where $\kappa \in]1, 2]$ is the TSE parameter of ψ. The subscript k underscores the fact that Ψ_k depends on a_{k-1}. Observe that Ψ_k is non-negative. The first part of Ψ_k corresponds to the potential energy of IBPG seen as a dissipative (discrete) dynamical system. The second term, which captures how the iterates remain close to each other, can be interpreted as a discrete Bregman version of the kinetic energy (involving the discrete velocity) of the system. The following lemma shows that Ψ_k is indeed a Lyapunov function for IBPG.

Lemma 5.6.2. *Under Assumptions 5.1.1 and (B.2), there exists $\nu \in]0, 1]$ such that the sequences generated by Algorithm 5 satisfy $\forall k \geq 0$*

$$\Psi_{k+1}(x_{k+1}, x_k) \leq \Psi_k(x_k, x_{k-1}) - \nu L D_\psi(x_k, x_{k-1}) - \left(\overline{a}^{1-\kappa} - 1\right) L D_\psi(x_{k+1}, y_k). \tag{5.6.1}$$

Proof. By L-smoothness of F relative to ψ, we have

$$\Phi(x_{k+1}) \leq F(y_k) + \langle \nabla F(y_k), x_{k+1} - y_k \rangle + G(x_{k+1}) + L D_\psi(x_{k+1}, y_k). \tag{5.6.2}$$

Observe that

$$x_{k+1} \in \operatorname*{Argmin}_{x \in \mathbb{R}^n} F(y_k) + \langle \nabla F(y_k), x - y_k \rangle + G(x) + \frac{1}{\gamma_k} D_\psi(x, y_k). \tag{5.6.3}$$

Then convexity of G gives that $\forall x \in \mathbb{R}^n$ (see *e.g.* [62, Lemma 3.2])

$$F(y_k) + \langle \nabla F(y_k), x_{k+1} - y_k \rangle + G(x_{k+1}) + \frac{1}{\gamma_k} D_\psi(x_{k+1}, y_k) \leq F(y_k) + \langle \nabla F(y_k), x - y_k \rangle + G(x) + \frac{1}{\gamma_k} D_\psi(x, y_k) - \frac{1}{\gamma_k} D_\psi(x_{k+1}, x). \tag{5.6.4}$$

Inserting this into (5.6.2), we get

$$\begin{aligned}\Phi(x_{k+1}) &\leq F(y_k) + \langle \nabla F(y_k), x - y_k\rangle + G(x) + \frac{1}{\gamma_k} D_\psi(x, y_k) - \frac{1}{\gamma_k} D_\psi(x_{k+1}, x) - \left(\frac{1}{\gamma_k} - L\right) D_\psi(x_{k+1}, y_k)\\ &= \Phi(x) - D_F(x, y_k) + \frac{1}{\gamma_k} D_\psi(x, y_k) - \frac{1}{\gamma_k} D_\psi(x_{k+1}, x) - \left(\frac{1}{\gamma_k} - L\right) D_\psi(x_{k+1}, y_k)\\ &\leq \Phi(x) + \left(\frac{1}{\gamma_k} + L\right) D_\psi(x, y_k) - \frac{1}{\gamma_k} D_\psi(x_{k+1}, x) - \left(\frac{1}{\gamma_k} - L\right) D_\psi(x_{k+1}, y_k),\end{aligned}$$

where we used again L-smoothness of F relative to ψ. Applying this inequality at $x = x_k$, we obtain

$$\Phi(x_{k+1}) + \frac{1}{\gamma_k} D_\psi(x_{k+1}, x_k) \leq \Phi(x_k) + \left(\frac{1}{\gamma_k} + L\right) D_\psi(x_k, y_k) - \left(\frac{1}{\gamma_k} - L\right) D_\psi(x_{k+1}, y_k).$$

We now use the TSP property twice to get

$$\begin{aligned}&D_\psi(x_k, y_k) = D_\psi((1-a_k)x_k + a_k x_k, (1-a_k)z_k + a_k x_k) \leq (1-a_k)^\kappa D_\psi(x_k, z_k) \text{ and}\\ &D_\psi(x_k, z_k) = D_\psi((1-a_{k-1})x_k + a_{k-1}x_k, (1-a_{k-1})x_{k-1} + a_{k-1}x_k) \leq (1-a_{k-1})^\kappa D_\psi(x_k, x_{k-1}).\end{aligned}$$

Combining the above inequalities, we arrive at

$$\begin{aligned}\Phi(x_{k+1}) + \frac{1}{\gamma_k} D_\psi(x_{k+1}, x_k) \leq \Phi(x_k) + \left(\frac{1}{\gamma_k} + L\right)(1-a_k)^\kappa (1-a_{k-1})^\kappa D_\psi(x_k, x_{k-1})\\ - \left(\frac{1}{\gamma_k} - L\right) D_\psi(x_{k+1}, y_k).\end{aligned}$$

Thus, in view of the choice of the parameters, there exists $\nu]0, 1]$ such that

$$\begin{aligned}\Phi(x_{k+1}) + a_k^{1-\kappa} L D_\psi(x_{k+1}, x_k) &\leq \Phi(x_k) + a_{k-1}^{1-\kappa} L D_\psi(x_k, x_{k-1}) - \nu a_{k-1}^{1-\kappa} L D_\psi(x_k, x_{k-1})\\ &\quad - \left(a_k^{1-\kappa} - 1\right) L D_\psi(x_{k+1}, y_k)\\ &\leq \Phi(x_k) + a_{k-1}^{1-\kappa} L D_\psi(x_k, x_{k-1}) - \nu L D_\psi(x_k, x_{k-1})\\ &\quad - \left(\overline{a}^{1-\kappa} - 1\right) L D_\psi(x_{k+1}, y_k).\end{aligned}$$

Subtracting $\inf \Phi$ on both sides, we get the claimed inequality. □

Capitalizing on the above descent property of the Lyapunov function Ψ_k, we get some preliminary convergence properties[2].

Proposition 5.6.3. *Under Assumptions 5.1.1 and (B.1)-(B.2), the sequences generated by Algorithm 5 (IBPG) are such that:*

(i) The sequence $\{\Psi_k(x_k, x_{k-1})\}_{k\in\mathbb{N}}$ *is decreasing and thus* $\lim_{k\to+\infty} \Psi_k(x_k, x_{k-1}) \geq 0$ *exists.*

(ii) $\sum_{k\in\mathbb{N}} \|x_k - x_{k-1}\|^2 < \infty$ *and thus* $\lim_{k\to\infty} \|x_k - x_{k-1}\| = 0$.

(iii) We have the rate

$$\min_{0\leq i\leq k} \|x_i - x_{i-1}\|^2 \leq \frac{2(\sigma_\psi \nu L)^{-1} \Psi_0(x_0, x_{-1})}{k+1}.$$

Proof. (i) The first claim comes from Lemma 5.6.2 and non-negativity of D_ψ thanks to convexity of ψ. The existence of the limit then follows as Ψ_k is non-negative and decreasing (recall that Φ is bounded from below, $a_k \geq \underline{a} > 0$ and $\kappa \in]0, 1]$).

(ii) We use σ_ψ-strong convexity of ψ and then sum (5.6.1) dropping the last non-negative term to get

$$\begin{aligned}\frac{\sigma_\psi \nu L}{2} \sum_{i=0}^{k} \|x_i - x_{i-1}\|^2 &\leq \nu L \sum_{i=0}^{k} D_\psi(x_i, x_{i-1})\\ &\leq \Psi_0(x_0, x_{-1}) - \Psi_k(x_k, x_{k-1}) \leq \Psi_0(x_0, x_{-1}).\end{aligned} \tag{5.6.5}$$

[2]Only convexity of ψ is needed for claim (i) to hold.

Passing to the limit as $k \to +\infty$ we get the summability claim.

(iii) We have

$$(k+1) \min_{0 \le i \le k} \|x_i - x_{i-1}\|^2 \le \sum_{i=0}^{k} \|x_i - x_{i-1}\|^2 .$$

Combining this with (5.6.5), we conclude.

□

To prove global convergence, it is sufficient to show that IBPG is a descent-like method according to Definition 2.4.3 and then to invoke Theorem 2.4.4. For this we need to a construct an appropriate function Ψ that verifies the conditions of Definition 2.4.3. The sequence of functions Ψ_k could do the job if a_k is taken fixed, say equal to $a \in [\underline{a}, \overline{a}]$, for all $k \ge K$, where K is arbitrarily large (see the discussion in Remark 5.2.4). We therefore consider the energy function

$$\Psi(x_k, x_{k-1}) = \begin{cases} \Psi_k(x_k, x_{k-1}) & \text{if} \quad k < K, \\ \Phi(x_k) - \inf \Phi + a^{1-\kappa} L D_\psi(x_k, x_{k-1}) & \text{otherwise.} \end{cases}$$

Observe first that Ψ is KL since both Φ and ψ are. The following proposition shows that the sequence generated by IBPG is a descent-like sequence for the new Lyapunov function Ψ.

Proposition 5.6.4. *Assume that Assumptions 5.1.1 and Assumptions 5.2.1 hold. Let $(x_k)_{k\in\mathbb{N}}$ be a bounded sequence generated by Algorithm 5. Then $(x_k)_{k\in\mathbb{N}}$ is a gradient-like descent sequence. Moreover the set of cluster points of $(x_k)_{k\in\mathbb{N}}$ is a nonempty compact set of* $\mathrm{crit}(\Phi)$.

Proof.

- *Sufficient decrease condition.* From Lemma 5.6.2 σ_ψ-strong convexity of ψ, we have $\forall k \in \mathbb{N}$

$$\begin{aligned} \Psi(x_{k+1}, x_k) &\le \Psi(x_k, x_{k-1}) - \nu L D_\psi(x_k, x_{k-1}) \\ &\le \Psi(x_k, x_{k-1}) - \frac{\sigma_\psi \nu L}{2} \|x_k - x_{k-1}\|^2 \end{aligned}$$

  which shows (C.1) in Definition 2.4.3.

- *Relative error condition.* We have to show (C.2) in Definition 2.4.3.
  ψ is C^2 hence $D_\psi(\cdot,\cdot)$ is C^1 jointly in its arguments. The sum rule of the limiting subdifferential applies in this case and tells us that for $k \ge K$, we have

$$\partial\Psi(x_{k+1}, x_k) = \begin{pmatrix} \nabla F(x_{k+1}) + \partial G(x_{k+1}) + a^{1-\kappa} L(\nabla\psi(x_{k+1}) - \nabla\psi(x_k)) \\ -a^{1-\kappa} L \nabla^2\psi(x_k)(x_{k+1} - x_k) \end{pmatrix}. \tag{5.6.6}$$

  From the update equation of x_{k+1} by IBPG, we have

$$\nabla\psi(y_k) - \gamma\nabla F(y_k) - \nabla\psi(x_{k+1}) \in \gamma\partial G(x_{k+1}), \tag{5.6.7}$$

  where $\gamma = a^{\kappa-1}/L$. Set $v_{k+1} \overset{\text{def}}{=} (v^1_{k+1}, v^2_{k+1})$ where

$$\begin{aligned} v^1_{k+1} &= (\nabla F(x_{k+1}) - \nabla F(y_k)) + a^{1-\kappa} L \left(\nabla\psi(y_k) - \nabla\psi(x_k)\right) \\ \text{and} \quad v^2_{k+1} &= -a^{1-\kappa} L \nabla^2\psi(x_k)(x_{k+1} - x_k). \end{aligned}$$

  In view of (5.6.6) and (5.6.7), we have

$$v_{k+1} \in \partial\Psi(x_{k+1}, x_k).$$

  We shall now bound $\|v_{k+1}\|^2 = \left\|v^1_{k+1}\right\|^2 + \left\|v^2_{k+1}\right\|^2$. Since $(x_k)_{k\in\mathbb{N}}$ is bounded and $\nabla\psi$ is Lipschitz continuous on any bounded subset by (B.3), there exists L_ψ such that

$$\left\|v^2_{k+1}\right\| \le a^{1-\kappa} L L_\Psi \|x_{k+1} - x_k\| \le \underline{a}^{1-\kappa} L L_\psi \|x_{k+1} - x_k\| .$$

Moreover, $(x_k)_{k\in\mathbb{N}}$ is also bounded by Lemma 5.6.1. Assumption (B.3) then entails that there exist $L_F > 0$ such that

$$\left\|v_{k+1}^1\right\| \leq L_F \|x_{k+1} - x_k\| + L_F \|x_k - y_k\| + \underline{a}^{1-\kappa} L L_\psi \|x_k - y_k\| .$$

where we used that $a \in]0, 1]$ and $\kappa \in]1, 2]$. Now, by definition of the iterates

$$x_k - y_k = x_k - z_k - a(x_k - z_k) = (1 - a)(x_k - x_{k-1} - a(x_k - x_{k-1})) = (1 - a)^2(x_k - x_{k-1}).$$

Therefore

$$\begin{aligned}\left\|v_{k+1}^1\right\| &\leq L_F \|x_{k+1} - x_k\| + (L_F + \underline{a}^{1-\kappa} L L_\psi)(1 - a)^2 \|x_k - x_{k-1}\| \\ &\leq L_F \|x_{k+1} - x_k\| + (L_F + \underline{a}^{1-\kappa} L L_\psi)(1 - \underline{a})^2 \|x_k - x_{k-1}\| .\end{aligned}$$

Taking $\rho_2 = L_F + (L_F + \underline{a}^{1-\kappa} L L_\psi)(1 - \underline{a})^2 + \underline{a}^{1-\kappa} L L_\psi$, we get the claim.

- *Continuity condition.* Since $(x_k)_{k\in\mathbb{N}}$ is bounded, its set of cluster points is a nonempty set. It is also compact set as the intersection of compact sets. Let us consider a subsequence $(x_{k_j})_{j\in\mathbb{N}}$ that converges to some limit x^*. From Proposition 5.6.3(i) and Lemma 5.6.1, we have that $\left(x_{k_j+1}\right)_{k\in\mathbb{N}}$ and $\left(y_{k_j}\right)_{k\in\mathbb{N}}$ converge to the same limit. Now arguing as in (5.6.3)-(5.6.4), we get

$$\begin{aligned}G(x_{k_j+1}) &\leq G(x^*) + \left\langle \nabla F(y_{k_j}), x^* - x_{k_j+1} \right\rangle - \frac{1}{\gamma} D_\psi(x_{k_j+1}, y_{k_j}) + \frac{1}{\gamma} D_\psi(x^*, y_{k_j}) - \frac{1}{\gamma} D_\psi(x_{k_j+1}, x^*) \\ &\leq G(x^*) + \left\langle \nabla F(y_{k_j}), x^* - x_{k_j+1} \right\rangle + \frac{1}{\gamma} D_\psi(x^*, y_{k_j}).\end{aligned}$$

Passing to the limit as $j \to +\infty$ and using continuity of D_ψ we get that

$$\limsup_{j\to+\infty} G(x_{k_j+1}) \leq G(x^*).$$

Combining this with continuity of F proves (C.3) in Definition 2.4.3.

Let $x_{k_j} \to x^*$. We argued above that $y_{k_j} \to x^*$ and $x_{k_j+1} \to x^*$. We then have from (5.6.7), continuity of $\nabla\psi$ and ∇F, and sequential closedness of ∇G that

$$-\nabla F(x^*) \in \partial G(x^*),$$

i.e. $x^* \in \mathrm{crit}(\Phi)$. □

Proof of Theorem 5.2.3

Proof. (i) This claim comes from Proposition 5.6.3.

(ii) This is a consequence of Proposition 5.6.4 and Theorem 2.4.4 after observing from (5.6.6) that $\mathrm{crit}(\Psi) = \{(x_\star, x_\star) : x_\star \in \mathrm{crit}(\Phi)\}$.

(iii) We now turn to proving convergence to a global minimizer. We introduce the following extended variable, $\widetilde{x_k} = (x_k, x_{k-1}) \in \mathbb{R}^n \times \mathbb{R}^n$ such that $\Psi(x_k, x_{k-1}) = \Psi(\widetilde{x_k})$ for all $k \in \mathbb{N}$. Let us choose radius $r > \rho_2 > 0$ such that $\eta < \rho_1(r - \rho_2)^2$. Let us suppose that the initial point x_0 is chosen such that the following conditions hold,

$$\Phi(x^\star) = \Psi(\widetilde{x^\star}) \leq \Psi(\widetilde{x_0}) < \Psi(\widetilde{x^\star}) + \eta = \Phi(x^\star) + \eta \tag{5.6.8}$$

$$\|x_0 - x^\star\| + 2\sqrt{\frac{\Psi(\widetilde{x_0}) - \Psi(\widetilde{x^\star})}{\rho_1}} + \frac{\rho_2}{\rho_1}\varphi\Big(\Psi(\widetilde{x_0}) - \Psi(\widetilde{x^\star})\Big) < \rho. \tag{5.6.9}$$

The condition (C.1) combined with (5.6.8) imply that for any $k \in \mathbb{N}, \Psi(\widetilde{x^\star}) \leq \Psi(\widetilde{x_{k+1}}) \leq \Psi(\widetilde{x_0}) < \Psi(\widetilde{x^\star}) + \eta$, and moreover

$$\|x_{k+1} - x_k\| \leq \sqrt{\frac{\Psi(\widetilde{x_{k+1}}) - \Psi(\widetilde{x_{k+2}})}{\rho_1}} \leq \sqrt{\frac{\Psi(\widetilde{x_{k+1}}) - \Psi(\widetilde{x^\star})}{\rho_1}}. \tag{5.6.10}$$

We deduce that if for any $k \in \mathbb{N}, x_k \in B(x^\star, \rho)$ then $x_{k+1} \in B(x^\star, r)$. Indeed, by the triangle inequality

$$\|x_{k+1} - x^\star\| \leq \|x_k - x^\star\| + \sqrt{\frac{\Psi(\widetilde{x_{k+1}}) - \Psi(\widetilde{x^\star})}{\rho_1}} = \rho + (r - \rho) = r. \tag{5.6.11}$$

It remains to show that $\forall k \in \mathbb{N}, x_k \in B(x^\star, \rho)$. We argue by induction. The triangle inequality gives

$$\|x_1 - x^\star\| \leq \|x_0 - x^\star\| + \sqrt{\frac{\Psi(\widetilde{x_1}) - \Psi(\widetilde{x^\star})}{\rho_1}} \leq \|x_0 - x^\star\| + \sqrt{\frac{\Psi(\widetilde{x_0}) - \Psi(\widetilde{x^\star})}{\rho_1}} < \rho,$$

which means that $x_1 \in B(x^\star, \rho)$. We also have

$$\|x_{k+1} - x^\star\| \leq \|z_0 - x^\star\| + 2\|x_1 - x_0\| + \sum_{j=1}^{k} \|x_k - x_{k-1}\|,$$

Standard arguments with the KL inequality show that

$$\sum_{i=l+1}^{k} \|x_k - x_{k-1}\| \leq \frac{\rho_2}{\rho_1} \varphi\Big(\Psi(x_{l+1}, x_l) - \Psi(\widetilde{x^\star})\Big). \tag{5.6.12}$$

Applying this bound (5.6.12) with $l = 0$ and combining with (5.6.9) yields

$$\|x_{k+1} - x^\star\| \leq \|x_0 - x^\star\| + 2\sqrt{\frac{\Psi(\widetilde{x_0}) - \Psi(\widetilde{x^\star})}{\rho_1}} + \frac{\rho_2}{\rho_1} \varphi\Big(\Psi(\widetilde{x_0}) - \Psi(\widetilde{x^\star})\Big) < \rho,$$

which implies that $x_{k+1} \in B(x^\star, \rho)$.
If we start close enough to $x^\star$ so that (5.6.8)-(5.6.9) holds the sequence $(x_k)_{k\in\mathbb{N}}$ will remain in the neighborhood $B(x^\star, \rho)$ and converges to a critical point, say x^*. Moreover $\Psi(x_k) \to \Phi(x^*) \geq \Phi(x^\star)$. Let us assume that $\Phi(x^*) > \Phi(x^\star)$. Since Φ has the KL property at $x^\star$ and thus

$$\varphi'\left(\Phi(x^*) - \Phi(x^\star)\right) \operatorname{dist}(0, \partial\Phi(x^*) \geq 1.$$

This is a contradiction since $\varphi'(s) > 0$ for $s \in]0, \eta[$ and $\operatorname{dist}(0, \partial\Phi(x^*)) = 0$ since x^* is a critical point. We conclude that x^* is indeed a global minimizer.

□

5.7 Proofs of Local Convergence

At this juncture, we pause to present the following Lemma, which forms the essence of our framework.

Lemma 5.7.1. ***(A Firmly nonexpansive map)*** *Let $G \in \Gamma_0(\mathbb{R}^n)$ and ψ be a strongly convex function with full domain then the following map $J \overset{\text{def}}{=} (\nabla\psi + \partial G)^{-1}$ and $\mathrm{Id} - J$ are firmly non-expansive.*

Let us observe that if ψ is not strongly convex but just convex, then from (5.7.1) one has that J is only monotone.

Proof. Let $x, y \in \mathbb{R}^n$ such that $p = J(x)$ and $q = J(y)$. Since $G \in \Gamma_0(\mathbb{R}^n)$ we have the following statements

$$D_G^{-\nabla\psi(p)+x}(q, p) \geq 0 \text{ and } D_G^{-\nabla\psi(q)+y}(p, q) \geq 0.$$

If we sum up both inequalities, we find that

$$D_G^{-\nabla\psi(p)+x}(q, p) + D_G^{-\nabla\psi(q)+y}(p, q) \geq 0 \iff \langle Jx - Jy, x - y\rangle \geq D_\psi(q, p) + D_\psi(p, q). \tag{5.7.1}$$

Therefore we get the desired result through strong convexity of ψ and [22, Proposition 4.4]. □

5.7.1 Proof of Lemma 5.3.3

Proof. We have $0 \in \partial\Phi(x_\star)$, since $G \in \mathrm{PSF}_{x_\star}(\mathcal{M}_{x_\star})$ and $F \in C^1$ thanks to [111, Corollay 4.7] (smooth perturbation of partly smooth functions), we have that $\Phi \in \mathrm{PSF}_{x_\star}(\mathcal{M}_{x_\star})$. From the global convergence Theorem 5.2.3, we have $x_k \xrightarrow{\Phi} x_\star$. Let us consider the iteration of Algorithm 5 and define

$$v_{k+1} = \frac{1}{\gamma_k}\left(-\nabla\psi(x_{k+1}) + \nabla\psi(y_k)\right) + \nabla F(x_{k+1}) - \nabla F(y_k),$$

we have that $\forall k \in \mathbb{N}, v_k \in \partial\Phi(x_k)$. Besides,

$$\|v_{k+1}\| \leq \frac{1}{\gamma_k}\|\nabla\psi(x_{k+1}) - \nabla\psi(y_k)\| + \|\nabla F(x_{k+1}) - \nabla F(y_k)\|$$

Since $(x_k)_{k\in\mathbb{N}}$ is assumed to be bounded and so thus $(y_k)_{k\in\mathbb{N}}$ by Lemma 5.6.1, we deduce from Assumption (B.3) that there exists a positive scalar $M_1, M_2 > 0$ such that

$$\begin{aligned}
\|v_{k+1}\| &\leq \left(\frac{M_1}{\gamma_k} + M_2\right)\|x_{k+1} - y_k\|, \\
&= \left(\frac{M_1}{\gamma_k} + M_2\right)\|x_{k+1} - (1-a_k)(1-a_{k-1})x_{k-1} - (1-a_k)a_{k-1}x_k - a_k x_k\|, \\
&\leq \left(\frac{M_1}{\gamma_k} + M_2\right)\left(\|x_{k+1} - x_k\| + (1-a_k)(1-a_{k-1})\|x_k - x_{k-1}\|\right), \\
&\leq \left(\frac{M_1}{\gamma_k} + M_2\right)\|x_{k+1} - x_k\| + \left(\frac{M_1}{\gamma_k} + M_2\right)(1-a_k)(1-a_{k-1})\|x_k - x_{k-1}\|,
\end{aligned}$$

Therefore, we obtained that for $k \to \infty$ we get that $a_k \to a \in [\underline{a}, \overline{a}]$ implying that $(\gamma_k)_{k\in\mathbb{N}}$ converges. The latter combined with the fact that $x_k \to x_\star$, we deduce that $\|v_k\| \to 0$ as $k \to \infty$. We conclude that x_k identifies $\mathcal{M}_{x_\star}$ thanks to [75, Proposition 10.12].

(i) If the active manifold $\mathcal{M}_{x^\star}$ is an affine subspace, then $\mathcal{M}_{x^\star} = x^\star + T_{x^\star}$ due to the normal sharpness property, and the claim follows immediately.

(ii) When G is locally polyhedral around $x^\star$, $\mathcal{M}_{x^\star}$ is an affine subspace, and the identification of $(y_k)_{k\in\mathbb{N}}$ and $(z_k)_{k\in\mathbb{N}}$ follows from (i). For the rest, it is sufficient to observe that, by polyhedrality, for any x in $\mathcal{M}_{x^\star}$ near $x^\star$, $\partial G(x) = \partial G(x^\star)$, combining Fact 2.5.4 and Fact 2.5.5, we arrive at the second conclusion.

□

The next Lemma gives the spectral properties of the matrices defined (5.3.1).

Lemma 5.7.2. *Under the Assumption 5.1.1-5.2.1, let $x_\star \in \mathrm{RI}(\Phi) \cap \mathrm{ND}(\Phi)$ such that F is locally C^2 around $x_\star$. Then for any stepsize $\gamma \in]0, 1/L]$, we have*

(i) H_F is symmetric positive definite with eigenvalues in $]\gamma\sigma\sigma_\psi, L\lambda_{\max}(\nabla^2\psi(x_\star))\gamma]$.

(ii) V has eigenvalues in

$$\left[\lambda_{\max}(\nabla^2\psi(x_\star))\left(1-\gamma L\right), \lambda_{\max}(\nabla^2\psi(x_\star)) - \gamma\sigma_\psi\sigma\right[,$$

hence $H_\psi^{-1}V$ has eigenvalues in

$$\left[1 - \gamma L\frac{\lambda_{\max}(\nabla^2\psi(x_\star))}{\sigma_\psi}, 1 - \gamma\sigma\frac{\sigma_\psi}{\lambda_{\max}(\nabla^2\psi(x_\star))}\right[.$$

(iii) WH_ψ has eigenvalues in $]0, q_\psi(x_\star)]$.

(iv) If either $x_\star \in \mathrm{ND}(\Psi)$ or $\mathcal{M}_{x_\star}$ is affine then WV has eigenvalues in $]-\Lambda, \Lambda[$ where we recall that $\Lambda = |q_\psi(x_\star) - \gamma\sigma|$ with $q_\psi(x_\star) = \frac{\lambda_{\max}(\nabla^2\psi(x_\star))}{\sigma_\psi}$.

Proof.

(i) Combining the fact that F is locally C^2 and $x_\star \in \mathrm{RI}(\Psi)$ we get that $\exists \sigma > 0, \forall h \in T_{x_\star}$,

$$\left\langle h; \nabla^2 F(x_\star) h \right\rangle \geq \left\langle h; \sigma \nabla^2 \psi(x_\star) h \right\rangle \geq \sigma \sigma_\psi \|h\|^2$$

where the last part comes from the strong convexity of ψ. This implies that $\lambda_{\min}(H_F) \geq \gamma \sigma \sigma_\psi$. Since F is $L-$smooth relative to ψ, $\forall h \in T_{x_\star}$,

$$\left\langle h; \nabla^2 F(x_\star) h \right\rangle \leq L \left\langle h; \nabla^2 \psi(x_\star) h \right\rangle \leq L \lambda_{\max}(\nabla^2 \psi(x_\star)) \|h\|^2 .$$

(ii) We have that $V = H_\psi - H_F$ and that H_ψ has eigenvalues in $[\sigma_\psi, \lambda_{\max}(\nabla^2 \psi(x_\star))]$. We combine this with the previous claim on H_F and the eigenvalues of the difference of two positive definite matrices.
For the second claim, let us observe that $H_\psi^{-1} V$ is similar to the following matrice $\mathrm{Id} - H_\psi^{-1/2} H_F H_\psi^{-1/2}$. Besides, we have that $H_\psi^{-1/2} H_F H_\psi^{-1/2}$ has eigenvalues in $\left[\gamma\sigma \frac{\sigma_\psi}{\lambda_{\max}(\nabla^2\psi(x_\star))}, \gamma L \frac{\lambda_{\max}(\nabla^2\psi(x_\star))}{\sigma_\psi}\right[$ thus, the difference of two positive definite matrices.

(iii) From Remark 5.3.4, we have that W has eigenvalues in $]0, 1/\sigma_\psi]$ which implies that WH_ψ has eigenvalues in $\left]0, \frac{\lambda_{\max}(\nabla^2\psi(x_\star))}{\sigma_\psi}\right]$.

(iv) Let us observe that $WV = W^{1/2}\left(W^{1/2} V W^{1/2}\right) W^{-1/2}$, thus WV is similar to $W^{1/2} V W^{1/2}$. We have

$$\begin{aligned}\left\|W^{1/2} V W^{1/2}\right\| &\leq \left\|W^{1/2}\right\| \|V\| \left\|W^{1/2}\right\| \leq \left|\lambda_{\max}(\nabla^2\psi(x_\star)) - \gamma\sigma_\psi\sigma\right| \left\|W^{1/2}\right\|^2 . \\ &= \left|\frac{\lambda_{\max}(\nabla^2\psi(x_\star))}{\sigma_\psi} - \gamma\sigma\right| ,\end{aligned}$$

where we used the fact that either $x_\star \in \mathrm{ND}(\Psi)$ or $\mathcal{M}_{x_\star}$ is affine holds then from Remark 5.3.4 we have $\|W\| \leq \frac{1}{\sigma_\psi}$.

□

Let us define the following matrices *i.e.*,

$$H_F^k \overset{\text{def}}{=} \gamma_k \mathrm{P}_{T_{x_\star}} \nabla^2 F(x_\star) \mathrm{P}_{T_{x_\star}}, \quad V^k \overset{\text{def}}{=} H_\psi - H_F^k, \qquad U^k \overset{\text{def}}{=} \gamma_k \nabla^2_{\mathcal{M}_{x_\star}} \Phi(x_\star) \mathrm{P}_{T_{x_\star}} - H_F. \tag{5.7.2}$$

To enhance readability, we introduce simplified notation for any $k \in \mathbb{N}$,

$$b_k = (1 - a_k) a_{k-1} + a_k \text{ and } c_k = (1 - a_k)(1 - a_{k-1}), \tag{5.7.3}$$

$$b = 2a - a^2 \text{ and } c = (1 - a)^2. \tag{5.7.4}$$

From this notation, we have this obvious Lemma.

Lemma 5.7.3. *Let us consider the sequences define in (5.7.3), If $a_k \to a$ as $k \to \infty$ then we have $b_k \to b$ and $c_k \to c$.*

Following the work of [118, Section B], we also define the matrices

$$M_1^k = \left[bW(V^k - V), \quad cW(V^k - V)\right], \quad M_2^k = \left[(b_k - b)WV^k, \quad (c_k - c)WV^k\right].$$

Therefore, we have the following proposition.

Proposition 5.7.4. *Under the same assumptions as Proposition 5.3.6, for k large enough we have*

$$\|y_k - x_\star\| = O(\|d_k\|), \|r_{k+1}\| = O(\|d_k\|), \|x_{k+1} - y_k\| = O(\|d_k\|), \tag{5.7.5}$$

$$\|\nabla F(y_k) - F(x_{k+1})\| = O\left(\|d_k\|\right), \quad \|\nabla\psi(y_k) - \nabla\psi(x_{k+1})\| = O\left(\|d_k\|\right), \tag{5.7.6}$$

and

$$\left\|W(U^k - U)(x_{k+1} - x_\star)\right\| = o(\|d_k\|), \left\|M_1^k d_k\right\| = o(\|d_k\|), \left\|M_2^k d_k\right\| = o(\|d_k\|). \tag{5.7.7}$$

Proof.

- From the definition of the sequences we have,

$$\begin{aligned}\|y_k - x_\star\| &= \|(1-a_k)z_k + a_k x_k - x_\star\|, \\ &= \|(1-a_k)(zk - x_\star) + a_k(x_k - x_\star)\|, \\ &= \|(1-a_k)(1-a_{k-1})r_{k-1} + ((1-a_k)a_{k-1} + a_k)\, r_k\|, \\ &\leq (1-a_k)(1-a_{k-1})\left(\|r_{k-1}\| + \|r_k\|\right), \\ &\leq \sqrt{2}\,\|d_k\|.\end{aligned}$$

- We recall that $r_{k+1} = x_{k+1} - x_\star$ thus,

$$\begin{aligned}\|r_{k+1}\| &= \left\|(\nabla\psi + \gamma_k \partial G)^{-1}(\nabla\psi(y_k) - \gamma_k \nabla F(y_k)) - (\nabla\psi + \gamma_k\partial G)^{-1}(\nabla\psi(x_\star) - \gamma_k\nabla F(x_\star))\right\|, \\ &\leq \|\nabla\psi(y_k) - \nabla\psi(x_\star)\| + \gamma_k \|\nabla F(y_k) - \nabla F(x_\star)\|, \\ &\leq (M_1 + M_2\gamma_k)\,\|y_k - x_\star\|, \\ &\leq (M_1 + M_2\gamma_k)\,\sqrt{2}\,\|d_k\|,\end{aligned}$$

  where we used the non-expansiveness of the mapping $(\nabla\psi + \gamma_k\partial G)^{-1}$ (see Lemma 5.7.1), the boundedness of the sequence and again Assumption (A.3).

- From Assumption (A.3), there exists $M_1 > 0$ large enough such that

$$\|\nabla\psi(y_k) - \nabla\psi(x_{k+1})\| \leq M_1\,\|x_{k+1} - y_k\| = O(\|d_k\|).$$

  Similarly, there exists M_2 large enough such that

$$\|\nabla F(y_k) - \nabla F(x_{k+1})\| \leq M_2\,\|y_k - x_{k+1}\| \leq M_2\sqrt{2}\,\|d_k\| = O(\|d_k\|).$$

- Let us now turn to the proof of (5.7.7), from the definition of Φ, we have that

$$\lim_{k\to\infty} \frac{\left\|W(U^k - U)r_{k+1}\right\|}{\|r_{k+1}\|} \leq \lim_{k\to\infty} |\gamma_k - \gamma|\,\|W\|\,\|\nabla_{\mathcal{M}_{x_\star}}\Phi(x_\star)\mathrm{P}_{T_{x_\star}}\| = 0, \tag{5.7.8}$$

  since $\gamma_k \to \gamma$ which means that $\left\|W(U^k - U)r_{k+1}\right\| = o(\|r_{k+1}\|) = o(\|d_k\|)$, where we have used (5.7.5).

- We have that,

$$\begin{aligned}\lim_{k\to\infty} \frac{\left\|M_1^k d_k\right\|}{\|d_k\|} &= \lim_{k\to\infty} \frac{\left\|bW(V^k - V)r_k + cW(V^k - V)r_{k-1}\right\|}{\|d_k\|} \\ &\leq \lim_{k\to\infty} \frac{\max(|b|,|c|)\,\|W\|\left\|V^k - V\right\|(\|r_k\| + \|r_{k+1}\|)}{\|d_k\|} \\ &\leq \lim_{k\to\infty} \frac{\max(|b|,|c|)\,\|W\|\,|\gamma_k - \gamma|\,\|\mathrm{P}_{T_{x_\star}}\nabla^2 F(x_\star)\mathrm{P}_{T_{x_\star}}\|\,\sqrt{2}\,\|d_k\|}{\|d_k\|} \\ &= \lim_{k\to\infty} \sqrt{2}|\gamma_k - \gamma|\max(|b|,|c|)\,\|W\|\left\|\mathrm{P}_{T_{x_\star}}\nabla^2 F(x_\star)\mathrm{P}_{T_{x_\star}}\right\| = 0,\end{aligned}$$

  as the term $\max(|b|,|c|)\,\|W\|\,\|\mathrm{P}_{T_{x_\star}}\nabla^2 F(x_\star)\mathrm{P}_{T_{x_\star}}\|$ is bounded since $\|W\|\,\|\mathrm{P}_{T_{x_\star}}\nabla^2 F(x_\star)\mathrm{P}_{T_{x_\star}}\| \leq L\lambda_{\max}(\nabla^2\psi(x_\star))^{-1}\,\|H_\psi\| \leq 1$ and $\max(|b|,|c|) \leq 1$.

- To finish the proof, we have

$$\begin{aligned}\lim_{k\to\infty} \frac{\left\|M_2^k d_k\right\|}{\|d_k\|} &= \lim_{k\to\infty} \frac{\left\|(b_k - b)WV^k r_k + (c_k - c)WV^k r_{k-1}\right\|}{\|d_k\|}, \\ &\leq \lim_{k\to\infty} \frac{\max\left(|b_k - b|, |c_k - c|\right)\|W\|\left\|V^k\right\|\sqrt{2}\,\|d_k\|}{\|d_k\|}, \\ &\leq \lim_{k\to\infty} \max\left(|b_k - b|, |c_k - c|\right)\|W\|\left\|V^k\right\|\sqrt{2} = 0,\end{aligned}$$

where we use (5.7.5), the fact that $\left\|V^k\right\| \to \|V\| < \infty$, the boundedness of $\|W\|$ combined with Lemma 5.7.3 to get the result.

□

5.7.2 Proof of Proposition 5.3.6

Proof.
Since $(z_k)_{k\in\mathbb{N}}$ is a sequence generated by Algorithm 5 converging to $x_\star \in \text{crit}\Phi$. From the finite identification Lemma 5.3.3, there exists $K \in \mathbb{N}$ such that x_k is close enough to $x_\star$ for $k \geq K$. Let $T_{x_{k+1}}, T_{x_\star}$ be their corresponding tangent spaces, and define $\tau_{k+1} : T_{x_\star} \to T_{x_{k+1}}$ the parallel translation along the unique geodesic joining x_{k+1} to $x_\star$. From the definition of the point x_{k+1} and the fact that $x_\star \in \text{crit}\Phi$ we have

$$\begin{aligned}\nabla\psi(y_k) - \nabla\psi(x_{k+1}) - \gamma_k\left(\nabla F(y_k) - \nabla F(x_{k+1})\right) &\in \gamma_k \partial\Phi(x_{k+1})\\ 0 &\in \gamma_k\partial\Phi(x_\star).\end{aligned}$$

We project now this inclusions over the tangents spaces $T_{x_{k+1}}$ and $T_{x_\star}$ respectively to get that

$$\begin{aligned}\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\left(\nabla\psi(y_k) - \nabla\psi(x_{k+1}) - \gamma_k\left(\nabla F(y_k) - \nabla F(x_{k+1})\right)\right) &= \gamma_k\tau_{k+1}^{-1}\nabla_{\mathcal{M}_{x_\star}}\Phi(x_{k+1})\\ 0 &= \gamma_k\nabla_{\mathcal{M}_{x_\star}}\Phi(x_\star)\end{aligned}$$

where we have used the Fact 2.5.5. After summing both lines and subtracting the value $\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\nabla\psi(x_\star)$ we have

$$\begin{aligned}&\gamma_k\left(\tau_{k+1}^{-1}\nabla_{\mathcal{M}_{x_\star}}\Phi(x_{k+1}) - \nabla_{\mathcal{M}_{x_\star}}\Phi(x_\star)\right)\\ &= \tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\left(\nabla\psi(y_k) - \nabla\psi(x_{k+1})\right) - \gamma_k\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\left(\nabla F(y_k) - \nabla F(x_{k+1})\right).\end{aligned} \tag{5.7.9}$$

We combined Lemma 2.5.1, (5.7.6) and Lemma 2.5.2 due to the local C^2-smoothness of ψ to obtain that

$$\begin{aligned}&\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\left(\nabla\psi(y_k) - \nabla\psi(x_{k+1})\right)\\
&= \mathrm{P}_{T_{x_\star}}\left(\nabla\psi(y_k) - \nabla\psi(x_{k+1})\right) + o(\|x_{k+1} - y_k\|),\\
&= \mathrm{P}_{T_{x_\star}}\left(\nabla\psi(y_k) - \nabla\psi(x_\star)\right) - \mathrm{P}_{T_{x_\star}}\left(\nabla\psi(x_{k+1}) - \nabla\psi(x_\star)\right) + o(\|d_k\|),\\
&= \mathrm{P}_{T_{x_\star}}\nabla^2\psi(x_\star)(y_k - x_\star) + o(\|y_k - x_\star\|) - \mathrm{P}_{T_{x_\star}}\nabla^2\psi(x_\star)(x_{k+1} - x_\star) + o(\|r_{k+1}\|) + o(\|d_k\|),\\
&= \mathrm{P}_{T_{x_\star}}\nabla^2\psi(x_\star)(y_k - x_\star) - \mathrm{P}_{T_{x_\star}}\nabla^2\psi(x_\star)(x_{k+1} - x_\star) + o(\|d_k\|),\\
&= \mathrm{P}_{T_{x_\star}}\nabla^2\psi(x_\star)\mathrm{P}_{T_{x_\star}}(y_k - x_\star) - \mathrm{P}_{T_{x_\star}}\nabla^2\psi(x_\star)\mathrm{P}_{T_{x_\star}}(x_{k+1} - x_\star) + o(\|d_k\|),\end{aligned} \tag{5.7.10}$$

where we have used (5.7.6), [120, Lemma 5.1] and the fact that $r_{k+1} = O(\|d_k\|)$. Using similar arguments, we have

$$\begin{aligned}&\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\left(\nabla F(y_k) - \nabla F(x_{k+1})\right)\\
&= \mathrm{P}_{T_{x_\star}}\left(\nabla F(y_k) - \nabla F(x_{k+1})\right) + o(\|d_k\|),\\
&= \mathrm{P}_{T_{x_\star}}\left(\nabla F(y_k) - \nabla F(x_\star)\right) - \mathrm{P}_{T_{x_\star}}\left(\nabla F(x_{k+1}) - \nabla\psi(x_\star)\right) + o(\|d_k\|),\\
&= \mathrm{P}_{T_{x_\star}}\nabla^2 F(x_\star)(y_k - x_\star) + o(\|y_k - x_\star\|) - \mathrm{P}_{T_{x_\star}}\nabla^2 F(x_\star)(x_{k+1} - x_\star) + o(\|x_{k+1} - x_\star\|) + o(\|d_k\|),\\
&= \mathrm{P}_{T_{x_\star}}\nabla^2 F(x_\star)(y_k - x_\star) - \mathrm{P}_{T_{x_\star}}\nabla^2 F(x_\star)(x_{k+1} - x_\star) + o(\|d_k\|),\\
&= \mathrm{P}_{T_{x_\star}}\nabla^2 F(x_\star)\mathrm{P}_{T_{x_\star}}(y_k - x_\star) - \mathrm{P}_{T_{x_\star}}\nabla^2 F(x_\star)\mathrm{P}_{T_{x_\star}}(x_{k+1} - x_\star) + o(\|d_k\|).\end{aligned} \tag{5.7.11}$$

Moreover, we have

$$\tau_{k+1}^{-1}\nabla_{\mathcal{M}_{x_\star}}\Phi(x_{k+1}) - \nabla_{\mathcal{M}_{x_\star}}\Phi(x_\star) = \nabla^2_{\mathcal{M}_{x_\star}}\Phi(x_\star)\mathrm{P}_{T_{x_\star}}\left(x_{k+1} - x_\star\right) + o(\|d_k\|). \tag{5.7.12}$$

We replace now the expressions (5.7.10), (5.7.11), (5.7.12), in (5.7.9), we obtain

$$\Big(\mathrm{P}_{T_{x_\star}}\nabla^2\psi(x_\star)\mathrm{P}_{T_{x_\star}} + \gamma_k\nabla^2_{\mathcal{M}_{x_\star}}\Phi(x_\star)\mathrm{P}_{T_{x_\star}} - \gamma_k\mathrm{P}_{T_{x_\star}}\nabla^2 F(x_\star)\mathrm{P}_{T_{x_\star}}\Big)(x_{k+1}-x_\star) = (H_\psi + U^k)(x_{k+1}-x_\star)$$
$$= \mathrm{P}_{T_{x_\star}}\nabla^2\psi(x_\star)\mathrm{P}_{T_{x_\star}}(y_k - x_\star) - \gamma_k\mathrm{P}_{T_{x_\star}}\nabla^2 F(x_\star)\mathrm{P}_{T_{x_\star}}(y_k - x_\star) + o(\|d_k\|). \tag{5.7.13}$$

We get by factorizing and replacing (5.7.2) that

$$(H_\psi + U^k)(x_{k+1}-x_\star) = V^k(y_k - x_\star) + o(\|d_k\|). \tag{5.7.14}$$

Now we replace the expressions of the inertial term in term of x_k, x_{k-1} to obtain that

$$(H_\psi + U^k)r_{k+1} = \left((1-a_k)a_{k-1} + a_k\right)V^k r_k + (1-a_k)(1-a_{k-1})V^k r_{k-1} + o(\|d_k\|), \tag{5.7.15}$$

We can write further

$$\begin{aligned}(H_\psi + U)\,r_{k+1} = &\left(U - U^k\right)r_{k+1} + \left((1-a_k)a_{k-1} + a_k\right)V^k r_k + (1-a_k)(1-a_{k-1})V^k r_{k-1}\\ &+ o(\|d_k\|),\end{aligned}$$

Thanks to Remark 5.3.4, it is possible to invert the matrice $U + H_\psi$ we have,

$$\begin{aligned}r_{k+1} = W\left(U - U^k\right)r_{k+1} + \left((1-a_k)a_{k-1} + a_k\right)WV^k r_k + (1-a_k)(1-a_{k-1})WV^k r_{k-1}\\ + o(\|Wd_k\|),\end{aligned}$$

Let us use the notation (5.7.3) with the estimates (5.7.7)

$$d_{k+1} = \left(M + \begin{bmatrix}M_1^k\\ 0\end{bmatrix} + \begin{bmatrix}M_2^k\\ 0\end{bmatrix}\right)d_k + o(\|d_k\|) = Md_k + o(\|d_k\|),$$

which concludes the proof. □

5.7.3 Proof of Proposition 5.3.8

Proof. We have that

$$\begin{aligned}M\begin{pmatrix}r_1\\ r_2\end{pmatrix} &= \begin{bmatrix}(2a-a^2)H_\psi^{-1}V & (a-1)^2H_\psi^{-1}V\\ \mathrm{Id} & 0\end{bmatrix}\begin{pmatrix}r_1\\ r_2\end{pmatrix}\\ &= \begin{pmatrix}(2a-a^2)H_\psi^{-1}Vr_1 + (a-1)^2H_\psi^{-1}Vr_2\\ r_1\end{pmatrix} = \varrho\begin{pmatrix}r_1\\ r_2\end{pmatrix},\end{aligned}$$

therefore $r_1 = \varrho r_2$, and insert it in the first identity to get that

$$\varrho^2 r_2 = (2a-a^2)H_\psi^{-1}V\varrho r_2 + (1-a)^2H_\psi^{-1}Vr_2$$

which means that

$$\left((2a-a^2)\varrho + (1-a)^2\right)H_\psi V r_2 = \varrho^2 r_2$$

thus there exists η such that $H_\psi^{-1}Vr_2 = \eta r_2$ moreover η satisfies the following equation

$$\varrho^2 - (2a-a^2)\varrho\eta - (1-a)^2\eta = 0. \tag{5.7.16}$$

The rest of the proof follows exactly the same step as the proof of [118, Proposition 4.7,Corollary 4.9] and we conclude with Lemma 5.7.2-(iv). □

5.8 Proof of the Escape Property

5.8.1 Proof of Theorem 5.4.1

Let us define the following mapping for $a \in [\underline{a}, \overline{a}]$.

$$\mathbf{T}(x_2, x_1) \overset{\text{def}}{=} \begin{bmatrix} \nabla\psi^{-1}\left(\nabla\psi(y(x_2, x_1)) - \gamma\nabla F(y(x_2, x_1))\right) \\ x_2 \end{bmatrix}, \tag{5.8.1}$$

where

$$y(x_2, x_1) = (2a - a^2)x_2 + (1-a)^2 x_1.$$

It is simple to see that $x_\star \in \mathrm{crit}\Phi$ if and only if $(x_\star, x_\star)$ is a fixed point of the operator $\mathbf{T}$. We have the following lemma which is an extension of the [83, Lemma A.2] to the inertial case.

Lemma 5.8.1. *Let $\mathbf{T}$ be defined as in (5.8.1) then,*

(a) For all $(x_2, x_1) \in \mathbb{R}^{2n}, \det D\mathbf{T}(x_2, x_1) \neq 0$,

(b) The set of strict saddle points is contained in the following set

$$U_{\mathbf{T}} \overset{\text{def}}{=} \left\{ (x_2, x_1) \in \mathbb{R}^{2n} : \mathbf{T}(x_2, x_1) = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}, \max_i |\lambda_i(D\mathbf{T}(x_1, x_2))| > 1 \right\}, \tag{5.8.2}$$

$$= \left\{ (x, x) \in \mathbb{R}^{2n} : \mathbf{T}(x, x) = \begin{bmatrix} x \\ x \end{bmatrix}, \max_i |\lambda_i(D\mathbf{T}(x, x))| > 1 \right\}. \tag{5.8.3}$$

Proof.

(a) Since ψ is a C^2 function, and thus $\nabla\psi$ is C^1, and as ψ is strongly convex, the inverse function theorem ensures that $(\nabla\psi)^{-1}$ is a local diffeomorphism. Moreover, F is C^2, therefore we define the following for simplicity

$$\mathbf{A}(x_2, x_1) \overset{\text{def}}{=} (2a - a^2)\nabla^2\psi^{-1}\left(\nabla\psi(y(x_2, x_1)) - \gamma\nabla F(y(x_1, x_2))\right) \left(\nabla^2\psi(y(x_2, x_1)) - \gamma\nabla^2 F(y(x_2, y_1))\right),$$

and

$$\mathbf{B}(x_2, x_1) \overset{\text{def}}{=} (1-a)^2\nabla^2\psi^{-1}\left(\nabla\psi(y(x_2, x_1)) - \gamma\nabla F(y(x_1, x_2))\right) \left(\nabla^2\psi(y(x_2, x_1)) - \gamma\nabla^2 F(y(x_2, y_1))\right),$$

then one can write that

$$D\mathbf{T}(x_2, x_1) = \begin{bmatrix} \mathbf{A}(x_2, x_1) & \mathbf{B}(x_2, x_1) \\ \mathrm{Id} & 0 \end{bmatrix}. \tag{5.8.4}$$

We deduce that for any (x_2, x_1) we have that $\det D\mathbf{T}(x_2, x_1) = \det(-\mathbf{B}(x_2, x_1))$. Therefore, it suffices to show that $\det(-\mathbf{B}(x_2, x_1)) \neq 0$ which hold since $\forall (x_2, x_1)$, $\nabla^2\psi(y(x_2, x_1)) - \gamma\nabla^2 F(y(x_2, x_1))$ is invertible. Indeed we have,

$$\nabla^2\psi(y(x_2, x_1)) - \gamma\nabla^2 F(y(x_2, x_1)) \succ (1 - \gamma L)\nabla^2\psi(y(x_2, x_1)) \succ 0,$$

where we have used the $L-$smooth adaptable property and the strong convexity of ψ.

(b) Let $x_\star$ be a strict saddle point therefore $(x_\star, x_\star)$ is a fixed point of $\mathbf{T}$. To have that $(x_\star, x_\star) \in U_{\mathbf{T}}$ it remains to show that $D\mathbf{T}$ has an eigenvalue of magnitude greater than 1. We have

$$D\mathbf{T}(x_\star, x_\star) = \begin{bmatrix} (2a - a^2)\left(\mathrm{Id} - \gamma\nabla^2\psi(x_\star)^{-1}\nabla^2 F(x_\star)\right), & (1-a)^2\left(\mathrm{Id} - \gamma\nabla^2\psi(x_\star)^{-1}\nabla^2 F(x_\star)\right) \\ \mathrm{Id} & 0 \end{bmatrix}.$$

Let us remark that $D\mathbf{T}(x_\star, x_\star)$ is similar to the following matrix

$$\widetilde{\mathbf{T}} = \begin{bmatrix} (2a-a^2)\left(\mathrm{Id} - \gamma H_\psi^{-1/2}\nabla^2 F(x_\star)H_\psi^{1/2}\right), & (1-a)^2\left(\mathrm{Id} - \gamma H_\psi^{-1/2}\nabla^2 F(x_\star)H_\psi^{1/2}\right) \\ \mathrm{Id} & 0 \end{bmatrix},$$

where we have applied the following transformation $\begin{bmatrix} H_\psi^{-1/2} & 0 \\ 0 & H_\psi^{-1/2} \end{bmatrix} D\mathbf{T}(x_\star, x_\star) \begin{bmatrix} H_\psi^{1/2} & 0 \\ 0 & H_\psi^{1/2} \end{bmatrix}$. For $\gamma < 1/L$, the symmetric matrix $\mathrm{Id} - \gamma H_\psi^{-1/2}\nabla^2 F(x_\star)H_\psi^{1/2}$ has an eigenvalue of magnitude greater than one,(see Lemma 3.5.2). Let us denote by v the eigenvector associated with this eigenvalue that we denote $\eta > 1$. We claim that the vector $\begin{bmatrix} v \\ 0 \end{bmatrix}$ is an eigenvector associated with the eigenvalue η of the matrix $\widetilde{\mathbf{T}}$. Indeed, we have

$$\widetilde{\mathbf{T}}\begin{bmatrix} v \\ 0 \end{bmatrix} = \begin{bmatrix} (2a-a^2)v\eta + (1-a)^2 v\eta \\ 0 \end{bmatrix} = \eta \begin{bmatrix} v \\ 0 \end{bmatrix},$$

which conclude the proof.

□

To show our Claim, we combine the global convergence result of Theorem 5.2.3-(ii) the previous Lemma and the center stable manifold theorem see [110, Corollary 1]. This allows us to say that the set

$$\left\{(z_0, z_{-1}) \in \mathbb{R}^{2n} : \lim_{k\to\infty} \mathbf{T}^k((z_0, z_{-1})) \in \mathrm{strisad}(\Phi)\right\}$$

has measure zero.

5.8.2 Proof of Lemma 5.4.3

Let us first observe that since $x_\star \in \mathrm{ND}(\Phi)$, using the same arguments as for the finite time identification Lemma 5.3.3 for any $x = (x_2, x_1) \in \mathbb{R}^{2n}$ near $(x_\star, x_\star)$, $\mathbf{T}(x_2, x_1) \in \widetilde{\mathcal{W}}_{x_\star}$. The subsequent portion of the proof faithfully adheres to the perturbation analysis used in [68, Theorem 4.1] with a minor adjustment to the inertial Bregman case.

Let $\widetilde{G} : \mathbb{R}^n \to \mathbb{R}$ be any C^2-smooth extension (representative) of G on the neighborhood of $x_\star$ in $\mathcal{W}_{x_\star}$, consider the following problem defining $P_1\mathbf{T}$ near $(x_\star, x_\star)$,

$$\min_{u\in\mathcal{W}_{x_\star}} \Upsilon(x,u) \overset{\text{def}}{=} \left\{F(y(x)) + \widetilde{G}(u) + \langle \nabla F(y(x)), u - y(x)\rangle + \frac{1}{\gamma} D_\psi(u, y(x))\right\}. \tag{$\widetilde{\mathcal{P}}_x$}$$

We can write the map $\mathbf{T}$ as

$$\mathbf{T}(x) = \begin{bmatrix} \min\limits_{u\in\mathcal{W}_{x_\star}} \Upsilon(x,u) \\ x \end{bmatrix},$$

To apply the perturbation result [160, Theorem 3.1] to Υ. We need a quadratic growth condition which we got using the fact that ψ is strongly convex and thus $u \mapsto \Upsilon(x,u)$ is also $\frac{\sigma_\psi}{\gamma}$-strongly convex. We also need to check a level-boundedness condition. This comes from a sufficient condition see [68, Lemma 2.4]. $u(x)$, the minimizer of $\Upsilon(z,\cdot)$ is a continuous map in near $(x_\star, x_\star)$ then we apply the perturbation analysis [160, Theorem 3.1] with the following Lagrangian function

$$\mathcal{L}(x,u,\lambda) \overset{\text{def}}{=} \Upsilon(x,u) + \langle W(u), \lambda\rangle,$$

where λ is the vector of Lagrange multipliers.

Since $u \longmapsto \Upsilon(x,u)$ is a strongly convex function and $u(x_\star,x_\star)=x_\star$ is the unique minimizer on $\mathcal{W}_{x_\star}$. Thus there exists optimal Lagrange multipliers $\bar{\lambda}$ such that

$$\nabla_u \mathcal{L}\left((x_\star,x_\star),x_\star,\bar{\lambda}\right)=\nabla\widetilde{G}(x_\star)+\nabla F(x_\star)+\sum_{i=1}^{n}\bar{\lambda}_i\nabla W_i(x_\star)=0.$$

From [160, Theorem 3.1], we get that locally for any x near $(x_\star,x_\star)$, $u(x)=P_1\mathbf{T}(x)$ is a C^1- smooth map and we deduce that for any $h=(h_1,h_2)\in\mathbb{R}^{2n}$ with $\|h\|\to 0$,

$$\langle \nabla P_1\mathbf{T}(x_\star,x_\star),h\rangle = \underset{v\in\mathcal{T}_{\mathcal{W}_{x_\star}}}{\operatorname{argmin}} 2\left\langle \nabla^2_{zu}\mathcal{L}\left((x_\star,x_\star),x_\star,\bar{\lambda}\right)v,h\right\rangle+\left\langle \nabla^2_{uu}\mathcal{L}\left((x_\star,x_\star),x_\star,\bar{\lambda}\right)v,v\right\rangle. \quad (5.8.5)$$

The expression of the Hessians of the Lagrangian is of the form $2n\times n$

$$\begin{aligned}\mathcal{H}_{zu}&\overset{\text{def}}{=}\nabla^2_{zu}\mathcal{L}\left((x_\star,x_\star),x_\star,\bar{\lambda}\right)\\&=\begin{bmatrix}(2-a)\left(\nabla^2F(x_\star)-\frac{1}{\gamma}\nabla^2\psi(x_\star)\right),\\(1-a)^2\left(\nabla^2F(x_\star)-\frac{1}{\gamma}\nabla^2\psi(x_\star)\right)\end{bmatrix},\end{aligned}$$

and

$$\mathcal{H}_{uu}\overset{\text{def}}{=}\nabla^2_{uu}\mathcal{L}\left((x_\star,x_\star),x_\star,\bar{\lambda}\right)=\nabla^2\widetilde{G}(x_\star)+\frac{1}{\gamma}\nabla^2\psi(x_\star)+\sum_{i=1}^{n}\bar{\lambda}_i\nabla^2_{uu}W_i(x_\star).$$

The minimization problem (5.8.5) is a quadratic form over the tangent space $\mathcal{T}_{\mathcal{W}_{x_\star}}$ thus the minimizer $\bar{v}$ must satisfy the following condition

$$\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}\mathcal{H}_{uu}\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}\bar{v}+\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}P_1\mathcal{H}_{zu}^{\top}\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}h_1+\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}P_2\mathcal{H}_{zu}^{\top}\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}h_2=0. \quad (5.8.6)$$

Let us denote $\widetilde{\mathcal{H}}_{uu}=\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}\mathcal{H}_{uu}\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}$, $\widetilde{\mathcal{H}}^2_{zu}=\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}P_1\mathcal{H}_{zu}^{\top}\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}$ and $\widetilde{\mathcal{H}}^1_{zu}=\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}P_2\mathcal{H}_{zu}^{\top}\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}$. (5.8.6) becomes

$$\widetilde{\mathcal{H}}_{uu}\bar{v}+\widetilde{\mathcal{H}}^2_{zu}h_1+\widetilde{\mathcal{H}}^1_{zu}h_2=0.$$

Since $x_\star$ is the unique minimizer of $\Upsilon\left((x_\star,x_\star),\cdot\right)$ and we also observe that $\Upsilon\left((x_\star,x_\star),\cdot\right)$ has the same active manifold $\mathcal{W}_{x_\star}$ as a sum of G and a smooth function this implies that $\Upsilon\left((x_\star,x_\star),\cdot\right)$ has at least a quadratic growth near $x_\star$. Therefore $\widetilde{\mathcal{H}}_{uu}$ is symmetric positive definite and invertible. Then we solve (5.8.5) to get that

$$\langle \nabla P_1\mathbf{T}\left((x_\star,x_\star)\right),h\rangle=-\widetilde{\mathcal{H}}_{uu}^{-1}\widetilde{\mathcal{H}}^2_{zu}h_1-\widetilde{\mathcal{H}}_{uu}^{-1}\widetilde{\mathcal{H}}^1_{zu}h_2,\quad \forall h_1,h_2\in\mathcal{T}_{\mathcal{W}_{x_\star}}.$$

At this point, we get that $\mathbf{T}$ is a C^1-smooth map. We immediately deduce that

$$\langle \nabla\mathbf{T}\left((x_\star,x_\star)\right),h\rangle=\begin{bmatrix}-\widetilde{\mathcal{H}}_{uu}^{-1}\widetilde{\mathcal{H}}^2_{zu}h_2-\widetilde{\mathcal{H}}_{uu}^{-1}\widetilde{\mathcal{H}}^1_{zu}h_1\\ h_1\end{bmatrix},\quad \forall h_1,h_2\in\mathcal{T}_{\mathcal{W}_{x_\star}}.$$

It remains now to show that $\nabla\mathbf{T}(x_\star,x_\star)$ has at least one eigenvalue greater than one. Let us first show that when $x_\star\in\mathrm{Actstrisad}(\Phi)$, $-\widetilde{\mathcal{H}}_{uu}^{-1}\widetilde{\mathcal{H}}^2_{zu}$ has an eigenvalue greater than one. Let η be an real eigenvalue associated to an eigenvector $v\in\mathcal{T}_{\mathcal{W}_{x_\star}}$ *i.e.*

$$-\widetilde{\mathcal{H}}_{uu}^{-1}\widetilde{\mathcal{H}}^2_{zu}v=\eta v\iff\left(\eta\widetilde{\mathcal{H}}_{uu}+\widetilde{\mathcal{H}}^2_{zu}\right)v=0.$$

Let us observe that since $x_\star$ is an active strict saddle point for the problem, then $\widetilde{\mathcal{H}}_{uu}+\widetilde{\mathcal{H}}^2_{zu}$ has a strict negative eigenvalue. Indeed,

$$\widetilde{\mathcal{H}}_{uu}+\widetilde{\mathcal{H}}^2_{zu}=\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}\left(\nabla^2\widetilde{G}(x_\star)+\sum_{i=1}^{n}\bar{\lambda}_i\nabla^2_{uu}W_i(x_\star)+(1-a)^2\left(\nabla^2F(x_\star)-\frac{1}{\gamma}\nabla^2\psi(x_\star)\right)\right)\mathrm{P}_{\mathcal{T}_{\mathcal{W}_{x_\star}}}.$$

Combining with the fact that $\widetilde{\mathcal{H}}_{uu}$ is positive definite means that there exists $\eta>1$ such that the matrix $\eta\widetilde{\mathcal{H}}_{uu}+\widetilde{\mathcal{H}}_{zu}$ is singular. By construction, we get that $\eta>1$ is an eigenvalue of $\nabla\mathbf{T}$ associated to the eigenvector $\begin{bmatrix}v\\0\end{bmatrix}$ of $\mathbf{T}$.

Chapter 6

Low Complexity Regularized Phase Retrieval

In this chapter, we study the phase retrieval problem in the situation where the vector to be recovered has an a priori structure that can encoded into a regularization term. This regularizer is intended to promote solutions conforming to some notion of simplicity or low complexity. We investigate both noiseless recovery and stability to noise and provide a very general and unified analysis framework that goes far beyond the sparse phase retrieval mostly considered in the literature. In the noiseless case we provide sufficient conditions under which exact recovery, up to global sign change, is possible. For (sub)Gaussian measurements, we also provide sample complexity bounds for exact recovery. The depends on the Gaussian width of the descent cone at the sought-after vector which is a geometric measure of the complexity of $\bar{x}$. In the noisy case, we consider both the constrained (Mozorov) and penalized (Tikhonov) formulations. We provide sufficient conditions for stable recovery and prove linear convergence for sufficiently small noise. For Gaussian measurements, we again provide sample complexity bounds for this to hold in high probability. These bound depend on the the intrinsic dimension of the sought-after vector and only (poly)logarithmically on the ambient dimension.

Our main contributions are as follows:

Main contributions of this chapter

- Analysis of exact recovery for the noiseless regularized phase retrieval ($\mathcal{P}_{\bar{y},0}$) in the deterministic case, and explicit sample complexity bounds for standard (sub)Gaussian sensing vectors over a large class of regularizers.
- Analysis of stable phase retrieval in the deterministic case with linear convergence in the low noise regime. We also provide sample complexity bounds ensuring local stable recovery from standard Gaussian measurements.
- Instantiation of the above sample complexity bounds for several examples including the ℓ_1-norm, the $\ell_{1,2}$-norm as well as total variation.

Contents

6.1 Introduction

6.1.1 Problem statement

We consider a generic additive noise model in which the noisy phase retrieval problem reads:

$$\begin{cases} \text{Recover } \bar{x} \in \mathbb{R}^n \text{ from the measurements } y \in \mathbb{R}^m \\ y[r] = |\langle a_r, \bar{x} \rangle|^2 + \epsilon[r], \quad r \in [\![m]\!], \end{cases} \tag{GeneralPR}$$

where $[r]$ is the r-th entry of the corresponding vector, and $\epsilon \in \mathbb{R}^m$ is the noise vector. This model is inspired by works such as [55, 70, 63]. Since $\bar{x}$ is real-valued, the best one can hope for is to ensure that $\bar{x}$ is uniquely determined from its intensities up to a global sign. Without any prior information, one can recover the signal in the noiseless case using mirror descent Chapter 3 and Chapter 4 in the noisy case. In order to reach the land of well-posedness without unreasonably increasing the number of measurements, it appears natural to restrict the inversion process to a well-chosen low dimensional subset of $\mathbb{R}^n$ containing the plausible solutions including $\bar{x}$; e.g. a linear space or a union of subspaces. A closely related procedure, that we will describe shortly, amounts to adopting a variational framework where the sought-after solutions are those where a prior penalty/regularization function is the smallest. It is then natural to leverage this low dimensional structure which will hopefully allow to minimize the number of measurements needed for recovery, and this is the most important as the measurement process might be expensive or can destroy the sample at hand. Here, we focus on the Tikhonov variational regularization:

$$\inf_{x \in \mathbb{R}^n} \{F_{y,\lambda}(x) \stackrel{\text{def}}{=} \left\| y - |Ax|^2 \right\|^2 + \lambda R(x)\}, \tag{$\mathcal{P}_{y,\lambda}$}$$

where $A = [a_1, \ldots, a_m]^\top$ and $R : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function which is intended to promote objects similar or close to $\bar{x}$. $\lambda > 0$ is the regularization parameter which balances the

trade-off between fidelity and regularization. It is immediate that $x \mapsto \left\| y - |Ax|^2 \right\|^2$ is $C^2(\mathbb{R}^n)$ but is nonconvex due to the quadratic measurements (though weakly convex). Besides, his gradient is not Lipschitz continuous. In this setting, we can associate to the objective the following function or kernel

$$\psi(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2 . \tag{6.1.1}$$

We bear in mind that $x \mapsto \left\| y - |Ax|^2 \right\|^2$ is smooth relative to ψ (see Chapter 3). Therefore $F_{y,\lambda}(x)$ is amenable to the efficient Bregman proximal gradient scheme or its inertial version studied in Chapter 5.

It is well known in the inverse problem literature, see e.g. [158], that the value of λ should typically be an increasing function of $\|\epsilon\|$. In the special case where there is no noise, i.e. $\epsilon = 0$, the fidelity to data should be perfect, which corresponds to considering the limit[1] of ($\mathcal{P}_{y,\lambda}$) as $\lambda \to 0^+$. This limit turns out to be the noiseless version for exact (up to a global sign) recovery,

$$\inf_{x\in\mathbb{R}^n} R(x) \quad \text{s.t.} \quad |Ax| = \sqrt{\bar{y}} \quad \text{where} \quad \bar{y} \overset{\text{def}}{=} |A\bar{x}|^2. \tag{$\mathcal{P}_{\bar{y},0}$}$$

Denoting $\overline{\mathcal{F}} \overset{\text{def}}{=} \{w \in \mathbb{R}^m : \ |w| = \sqrt{\bar{y}}\}$, which is a non-empty finite bounded set of cardinality 2^m (vertices of a hyper-rectangle), ($\mathcal{P}_{\bar{y},0}$) can be equivalently written as

$$\inf_{x\in\mathbb{R}^n} R(x) \quad \text{s.t.} \quad Ax \in \overline{\mathcal{F}}.$$

We refer to Section 1.2.2 for an extended review about this problem.

6.1.2 Contributions

In this chapter, we start by providing sufficient conditions under which the set of solutions to ($\mathcal{P}_{\bar{y},0}$) is non-empty. Then, we show that the recovery of $\bar{x}$ up to a global sign is exact when we solve ($\mathcal{P}_{\bar{y},0}$) under two geometric (deterministic) conditions on R, the descent cone of R and the deterministic measurements A. It turns out that for standard Gaussian measurements and the class of regularizer that we consider, these conditions are satisfied with high probability under a sufficiently large sample complexity. As a consequence, when the number of measurements is large enough the recovery of $\bar{x}$ up to a sign change is exact by solving ($\mathcal{P}_{\bar{y},0}$). Futhermore, we provide an explicit expression of the recovery bounds for decomposable regularizers (including the lasso, the group lasso), for frame analysis-type regularizers and the total variation.

Concerning stable recovery, we first consider a relaxed inequality constrained form ($\mathcal{P}_{y,\rho}$) which is known as the residual method or Mozorov formulation. We show that under the previous deterministic conditions, the set of solutions is nonempty. Moreover, the solutions are located in a ball of center $\bar{x}$ up to a sign-change and radius equal to the signal-to-noise ratio. For standard Gaussian measurements and a large class of regularizers, we show with high probability that solving ($\mathcal{P}_{y,\rho}$) yields a solution that is near $\bar{x}$ up to a sign change as soon as the number of measurements is large enough.

We then turn to penalized problem ($\mathcal{P}_{y,\lambda}$). First, we show that under an appropriate geometric deterministic, the problem has a nonempty and compact set of minimizers. Then, using $\Gamma-$convergence tools, when $\lambda \to 0$ and $\epsilon \to 0$ the set of minimizers reduces to the set of true vectors up to a global sign change. Finally, we show that for small noise, the recovery error scales as $\|\epsilon\|$, a rate known in the inverse problem literature as linear convergence[2].

For standard Gaussian measurements, we exemplify our sample complexity bounds for several regularizers. This covers both the popular sparse retrieval case, but we also provide bounds that are new and unknown in the literature to the best of our knowledge.

[1] This will be studied rigorously in Section 6.4.

[2] The reason is that the bound is indeed linear $\|\epsilon\|$.

6.2 Noiseless Recovery

We here study well-posedness (existence and uniqueness of minimizers) of ($\mathcal{P}_{\bar{y},0}$), which in turn will allow us to state when exact recovery is possible. In this section, we use the shorthand notation

$$\mathcal{S}_{\bar{y},0} \stackrel{\text{def}}{=} \operatorname*{Argmin}_{A^{-1}(\overline{\mathcal{F}})}(R).$$

6.2.1 Existence of minimizers

The following result provides sufficient conditions under which problem ($\mathcal{P}_{\bar{y},0}$) has minimizers. It does not need convexity of R.

Proposition 6.2.1. *Let $R : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper and lsc function. Assume that:*

(i) $A(\operatorname{dom}(R)) \cap \overline{\mathcal{F}} \neq \emptyset$.

(ii) R *is non-negative*[3].

(iii) $\ker(R_\infty) \cap \ker(A) = \{0\}$.

Then $\mathcal{S}_{\bar{y},0}$ is a non-empty compact set.

Remark 6.2.2.

- A typical case where all above assumptions are in force is when R is coercive, has full domain and is bounded from below.
- This result is general and goes beyond the phase retrieval problem, indeed this result can be applied for instance for general non-linear inverse problem with a suitable definition of $\overline{\mathcal{F}}$.

Proof. The range of R_∞ is on $\mathbb{R}_+$ since R verifies (ii). Define $G = R + \iota_{\overline{\mathcal{F}}} \circ A$. In view of the domain qualification assumption (i), we get by [12, Proposition 2.6.1 and Proposition 2.6.3] that

$$G_\infty(z) \geq R_\infty(z) + \iota_{\overline{\mathcal{F}}_\infty}(Az).$$

Since $\overline{\mathcal{F}}$ is bounded, we get that $\overline{\mathcal{F}}_\infty = \{0\}$. Moreover, the range of R_∞ is on $\mathbb{R}_+$ since R is bounded from below. Thus

$$G_\infty(z) > 0 \quad \text{for all} \quad z \notin \ker(R_\infty) \cap \ker(A).$$

It then follows from [12, Corollary 3.1.2] that (iii) entails the claim. □

6.2.2 Deterministic recovery condition

Definition 6.2.3 (Descent cone). The descent cone of R at $\bar{x}$ is the conical hull of the sublevel set of R at $\bar{x}$, i.e.

$$\mathcal{D}_R(\bar{x}) \stackrel{\text{def}}{=} \bigcup_{t>0} \{z : \; R(\bar{x} + tz) \leq R(\bar{x})\}. \tag{6.2.1}$$

The tangent cone of the sublevel set of R at $\bar{x}$, denoted $\mathcal{T}_R(\bar{x}) \stackrel{\text{def}}{=} \overline{\operatorname{cone}}(\mathcal{S}_{\operatorname{lev}R}(\bar{x}) - \bar{x})$, is the closure of $\mathcal{D}_R(\bar{x})$. The normal cone of the sublevel set of R at $\bar{x}$ is

$$\mathcal{N}_R(\bar{x}) \stackrel{\text{def}}{=} \{s : \langle s, z - \bar{x}\rangle \leq 0, z \in \mathcal{S}_{\operatorname{lev}R}(\bar{x})\},$$

and we have $\mathcal{N}_R(\bar{x}) = \mathcal{T}_R(\bar{x})^\circ$, where $^\circ$ stand for polarity.

Theorem 6.2.4. *Suppose that $\mathcal{S}_{\bar{y},0} \neq \emptyset$, and that:*

(H.1) $R \in \Gamma_0(\mathbb{R}^n)$ *and is even symmetric.*

[3] In fact, we need R to be only bounded from below, and there is no loss of generality by taking the lower bound as 0 by a trivial translation argument.

(**H.2**) $\forall I \subset [\![m]\!], |I| \geq m/2$

$$\ker(A^I) \cap \mathcal{D}_R(\bar{x}) = \{0\}.$$

Then the recovery of $\bar{x}$ (up to a global sign) is exact by solving ($\mathcal{P}_{\bar{y},0}$), i.e.

$$\mathcal{S}_{\bar{y},0} = \overline{\mathcal{X}}.$$

Remark 6.2.5.

- Assumption is quite general in the choice of the regularizer R it encompasses any convex atomic norm, or norms such as $\ell_1, \ell_{1,2}$ and $\ell_\infty-$norm *etc.*
- Of course, assumption (**H.2**) is vacuous if $\mathcal{D}_R(\bar{x})$ is empty, which is the case if the set of minimizers is empty. The assumptions of Proposition 6.2.1 ensure that this cannot be the case.

Proof. The proof is a generalization of that [178, Theorem 2.2] beyond the ℓ_1-norm, and exploits the structure of the constraint set $\overline{\mathcal{F}}$. Let $b \stackrel{\text{def}}{=} A\bar{x}$, and for any sign vector $\varepsilon \in \{1,-1\}^m$, set $b_\varepsilon \stackrel{\text{def}}{=} [\varepsilon[r]b[r] : r \in [\![m]\!]]^\top$. Consider the minimization problem

$$\min_{x\in\mathbb{R}^n} R(x) \quad \text{s.t.} \quad Ax = b_\varepsilon,$$

and denote x_ε any minimizer, if it exists. If x_ε does not exist, there is nothing to say. We claim that if x_ε exists, then under our assumptions, for any sign vector ε,

$$R(\bar{x}) \leq R(x_\varepsilon)$$

with equality iff $x_\varepsilon = \pm\bar{x}$.

Observe that $x_\varepsilon \in A^{-1}(\overline{\mathcal{F}})$. Thus $\langle a_r, x_\varepsilon\rangle = \pm b[r]$ for all $r \in [\![m]\!]$. Let

$$I = \{r \in [\![m]\!] : \langle a_r, x_\varepsilon\rangle = b[r]\}.$$

Thus either $|I| \geq m/2$ or $|I^c| \geq m/2$. Assume the first case holds. This implies that $A^I x_\varepsilon = A^I \bar{x}$. From [61, Proposition 2.1], it follows using (**H.2**) and convexity of R that

$$\operatorname*{Argmin}_{x\in\mathbb{R}^n}\{R(x) \text{ s.t. } A^I x = A^I\bar{x}\} = \{\bar{x}\},$$

and thus, since x_ε is a feasible point,

$$R(\bar{x}) \leq R(x_\varepsilon),$$

with equality holding if and only if $x_\varepsilon = \bar{x}$. For the case where $|I^c| \geq m/2$, we have $-A^{I^c}x_\varepsilon = A^{I^c}\bar{x}$. Arguing similarly as before using also that R is even, we get

$$\operatorname*{Argmin}_{x\in\mathbb{R}^n}\{R(x) \text{ s.t. } -A^{I^c}x = A^{I^c}\bar{x}\} = -\operatorname*{Argmin}_{x\in\mathbb{R}^n}\{R(x) \text{ s.t. } A^{I^c}x = A^{I^c}\bar{x}\} = \{-\bar{x}\},$$

Thus, in this case

$$R(\bar{x}) \leq R(x_\varepsilon),$$

with equality holding if and only if $x_\varepsilon = -\bar{x}$. Since this holds for any $\varepsilon \in \{1,-1\}^m$ and any minimizer of ($\mathcal{P}_{\bar{y},0}$) is of the form x_ε (when the latter exists), we conclude. □

6.2.3 Recovery from Gaussian measurements

Here the entries of A are i.i.d. $\mathcal{N}(0, 1/m)$.

Lemma 6.2.6. *Let $\delta \in]0,1[$ and $\nu = \frac{1}{18}\sqrt{\frac{\pi}{2}}$. Suppose that $x \in \mathbb{R}^n$ is a fixed vector. Then*

$$\min_{I\subset[\![m]\!],|I|\geq m/2} \left\|A^I x\right\| \geq \nu/2\,\|x\|$$

with probability at least $1 - 2e^{-\frac{\nu^2 m}{8}}$, and

$$\max_{I\subset[\![m]\!],|I|\geq m/2} \left\|A^I x\right\| \leq (1+\delta)\,\|x\|$$

with probability at least $1 - e^{-\frac{\delta^2 m}{2}}$.

Proof. The first claim follows from [178, Lemma 4.4]. The second one follows from the fact that

$$\left\|A^I x\right\| \leq \|Ax\| \quad \text{for all} \quad I \subset [\![m]\!],$$

and then use Proposition 2.6.8. □

Theorem 6.2.7. *Let* ν *be as defined in Lemma 6.2.6. Suppose that* (**H.1**) *holds. Let* A *be a matrix whose entries are i.i.d.* $\mathcal{N}(0, 1/m)$ *with*

$$m \geq \frac{16}{\nu^2} \log\left(N\left(\mathcal{D}_R(\bar{x}) \cap \mathbb{S}^{n-1}, \varepsilon\right)\right),$$

for some $\varepsilon \in]0, \nu/(2+\nu)[$. *Then with probability at least* $1 - 3e^{-\frac{\nu^2 m}{16}}$, *the recovery of* $\bar{x}$ *(up to a global sign) is exact by solving* $(\mathcal{P}_{\bar{y},0})$.

Proof. The proof relies on combining Theorem 6.2.4 and Lemma 6.2.6 together with a covering argument. Throughout the proof, denote $\Omega = \mathcal{D}_R(\bar{x}) \cap \mathbb{S}^{n-1}$. In view of Theorem 6.2.4, we need to prove that there exists $c \in]0, 1[$ such that

$$\min_{I \subset [\![m]\!], |I| \geq m/2} \left\|A^I z\right\| \geq c$$

for all $z \in \Omega$. Let $\Omega_\varepsilon = \{z_i : i \in [\![N(\Omega, \varepsilon)]\!]\}$ be an ε-net of Ω. For a fixed $z_i \in \Omega_\varepsilon$, Lemma 6.2.6 tells us that

$$\left\|A^I z_i\right\| \geq \nu/2$$

with probability at least $1 - 2e^{-\frac{\nu^2 m}{8}}$. Now, for an arbitrary but fixed $z \in \Omega$, there exists $z_j \in \Omega_\varepsilon$ such that $\|z - z_j\| \leq \varepsilon$. Thus

$$\min_{I \subset [\![m]\!], |I| \geq m/2} \left\|A^I z\right\| \geq \min_{I \subset [\![m]\!], |I| \geq m/2} \left\|A^I z_j\right\| - \max_{I \subset [\![m]\!], |I| \geq m/2} \left\|A^I (z - z_j)\right\| \geq \frac{\nu}{2} - \left(1 + \frac{\nu}{2}\right)\varepsilon$$

with probability at least $1 - 3e^{-\frac{\nu^2 m}{8}}$, where we took $\delta = \nu/2$ in Lemma 6.2.6 for the second inequality. Taking ε small as devised and setting $c = \frac{\nu}{2} - (1 + \frac{\nu}{2})\,\varepsilon \in]0, \nu/2[$, we deduce that

$$\min_{I \subset [\![m]\!], |I| \geq m/2} \left\|A^I z\right\| \geq c$$

holds for all $z \in \Omega$ with probability at least $1 - 3e^{\log(N(\Omega,\varepsilon)) - \frac{\nu^2 m}{8}}$. The bound on the number of measurements then leads to the claim. □

Estimating covering numbers is difficult to compute for general convex cones. On the other hand in [61, 5, 172], the authors developed a general recipe for estimating Gaussian widths of the descent cone (restricted to the unit sphere). This is the motivation behind the following corollary.

Corollary 6.2.8. *Let* ν *be as defined in Lemma 6.2.6. Suppose that* (**H.1**) *holds. Let* A *be a matrix whose entries are i.i.d.* $\mathcal{N}(0, 1/m)$ *with*

$$m \geq \frac{32(\nu+2)^2}{\nu^4} w\left(\mathcal{D}_R(\bar{x}) \cap \mathbb{S}^{n-1}\right)^2.$$

Then with probability at least $1 - 3e^{-\frac{\nu^2 m}{16}}$, *the recovery of* $\bar{x}$ *(up to a global sign) is exact by solving* $(\mathcal{P}_{\bar{y},0})$.

Proof. Use the lower bound of Proposition 2.6.3 and choose $\varepsilon = \frac{\nu}{\sqrt{2}(2+\nu)}$ in Theorem 6.2.7. □

Remark 6.2.9.

(i) Clearly, this result shows that the sample complexity bound for exact phase recovery by solving ($\mathcal{P}_{\bar{y},0}$) is nearly (up to constants) the same as for exact recovery from linear Gaussian measurements [61, 5]. However, one has to keep in mind that ($\mathcal{P}_{\bar{y},0}$) contains a non-convex constraint and the recovery results we have are not for an algorithmic scheme.

(ii) This result can not be extended to the general case of the Subgaussian random variable. The most technical reason is that it heavily relies on the Gaussian structure of the measurements which is related to the strong RIP property defined in [178]. For instance, [178, Remark 2.3] highlights the fact that basic subgaussian sensing such as Bernoulli ensemble which is defined as

$$\Pr(a_{j,i} = 1) = \Pr(a_{j,i} = -1) = \frac{1}{2},$$

does not satisfy this property. Let us observe that regardless of the number of measurements taken reconstruction of an element of the standard basis of $\mathbb{R}^n$ is not possible with from this measurement.

In the numerical section, we will report experimental results with a Bregman Proximal Gradient algorithm showing good empirical performance.

6.2.4 Recovery bounds for decomposable regularizers

We start by defining some essential geometrical objects that were introduced in [173].

Definition 6.2.10 (Model Subspace). Let $x \in \mathbb{R}^n$. We denote by e_x as

$$e_x = \mathrm{P}_{\mathrm{aff}(\partial R(x))}(0).$$

We denote

$$S_x = \mathrm{par}(\partial R(x)) \text{ and } T_x = S_x^{\perp}.$$

T_x is coined the *model subspace* of x associated to J.

It can be shown, see [173, Proposition 5], that $x \in T_x$, hence the name model subspace. When R is differentiable at x, we have $e_x = \nabla R(x)$ and $T_x = \mathbb{R}^n$. When R is the ℓ_1-norm (Lasso), the vector e_x is nothing but the sign of x. Thus, e_x can be viewed as a generalization of the sign vector. Observe also that $e_x = \mathrm{P}_{T_x}(\partial R(x))$, and thus $e_x \in T_x \cap \mathrm{aff}(\partial R(x))$. However, in general, $e_x \notin \partial R(x)$.

In this subsection, we will assume that R is a strong gauge in the sense of [173, Definition 6].

Definition 6.2.11 (Strong Gauge). R is a strong gauge if

$$R = \gamma_{\mathcal{C}}, \tag{6.2.2}$$

where $\mathcal{C}$ is a non-empty convex compact set containing the origin as an interior point, and $e_x \in \mathrm{ri}(\partial R(x))$.

Strong gauges have a nice decomposable description of $\partial R(x)$ in terms of e_x, T_x, S_x and $\sigma_{\mathcal{C}}$. More precisely, piecing together [173, Theorem 1, Proposition 4 and Proposition 5(iii)], we have

$$\partial R(x) = \mathrm{aff}(\partial R(x)) \cap \mathcal{C}^{\circ} = \{v \in \mathbb{R}^n : \ v_{T_x} = e_x \text{ and } \sigma_{\mathcal{C}}(v_{S_x}) \leq 1\}. \tag{6.2.3}$$

The Lasso, group Lasso, and nuclear norms are typical popular examples of (symmetric) strong gauges. Let us observe that strong symmetric gauges not only conform to **(H.1)** but also meet the requirements outlined in Proposition 6.2.1.

The following Lemma is a characterization of the Gaussian width of the descent cone of strong gauge function.

Lemma 6.2.12. *If R is a strong gauge of $\mathcal{C}$, then for any $x \in \mathbb{R}^n \setminus \{0\}$*

$$w\left(\mathcal{D}_R(x) \cap \mathbb{S}^{n-1}\right)^2 \leq \mathbb{E}\left(\sigma_{\mathcal{C}}(g_{S_x})^2\right) \|e_x\|^2 + \dim\left(T_x\right). \tag{6.2.4}$$

Proof. From [61, Proposition 3.6] which is an expression of the Gaussian width of a cone in terms of the dual of the cone, we have

$$w\left(\mathcal{D}_R(x) \cap \mathbb{S}^{n-1}\right)^2 \leq \mathbb{E}\left(\operatorname{dist}\left(g, \mathcal{D}_R(x)^\circ\right)^2\right) = \mathbb{E}\left(\operatorname{dist}\left(g, \mathcal{N}_R(x)\right)^2\right).$$

R being a strong gauge implies R is convex and has full domain, and thus ∂R is non-empty convex and compact valued at any $x \in \mathbb{R}^n$. Moreover, $\operatorname{Argmin}(R) = \{0\}$. It then follows from [155, Theorem 23.7] that for any $x \neq 0$

$$\mathcal{N}_R(x) = \bigcup_{t \geq 0} t\partial R(x),$$

where $t\partial R(x)$ is the dilation of the subdifferential through the scaling factor t. In turn, we get

$$w\left(\mathcal{D}_R(x) \cap \mathbb{S}^{n-1}\right)^2 \leq \mathbb{E}\left(\operatorname{dist}\left(g, \cup_{t\geq 0} t\partial R(x)\right)^2\right) \leq \inf_{t \geq 0} \mathbb{E}\left(\operatorname{dist}\left(g, t\partial R(x)\right)^2\right) \leq \mathbb{E}\left(\operatorname{dist}\left(g, \tilde{t}\partial R(x)\right)^2\right)$$

for any $\tilde{t} \geq 0$. Observe that in view of definition (6.2.3), we have

$$t\partial R(x) = \left\{v \in \mathbb{R}^n : \; v_{T_x} = t e_x \;\text{ and }\; \sigma_{\mathcal{C}}(v_{S_x}) \leq t\right\}. \tag{6.2.5}$$

We will now device an appropriate choice of $\tilde{t}$ and of a subgradient in $\partial R(x)$[4]. Let v be a random vcetor such that $v_{S_x} = g_{S_x}$ and $v_{T_x} = \sigma_{\mathcal{C}}(g_{S_x}) e_x$. Obviously, $v \in \sigma_{\mathcal{C}}(g_{S_x})\partial R(x)$ by (6.2.5). Thus

$$\begin{aligned}
w\left(\mathcal{D}_R(x) \cap \mathbb{S}^{n-1}\right)^2 &\leq \mathbb{E}\left(\|g - v\|^2\right) \\
&= \mathbb{E}\left(\|(g_{T_x} - v_{T_x}) + (g_{S_x} - v_{S_x})\|^2\right) \\
&= \mathbb{E}\left(\|g_{T_x} - \sigma_{\mathcal{C}}(g_{S_x}) e_x\|^2\right) \\
&\leq \mathbb{E}\left(\sigma_{\mathcal{C}}(g_{S_x})^2\right) \|e_x\|^2 + \mathbb{E}\left(\|g_{T_x}\|^2\right) \\
&= \mathbb{E}\left(\sigma_{\mathcal{C}}(g_{S_x})^2\right) \|e_x\|^2 + \dim(T_x),
\end{aligned}$$

where we used orthogonality of T_x and S_x in the first equality, independence of g_{T_x} and $\sigma_{\mathcal{C}}(g_{S_x})$ in third equality since g is Gaussian, and $\mathbb{E}\left(\|g_{T_x}\|^2\right) = \operatorname{tr}(\mathrm{P}_{T_x}) = \dim(T_x)$ in the last equality. □

ℓ_1 regularization ℓ_1 regularization (a.k.a. Lasso) is used to promote the sparsity of the minimizers, see [47] for a comprehensive review. It corresponds to choosing R as the ℓ_1-norm

$$R(x) = \|x\|_1 = \sum_{i=1}^n |x[i]|. \tag{6.2.6}$$

It is also referred to as ℓ_1-synthesis in the signal processing community.

We denote $(a_i)_{1 \leq i \leq n}$ the canonical basis of $\mathbb{R}^n$ and $\operatorname{supp}(x) \overset{\text{def}}{=} \{i \in [\![n]\!] : \; x[i] \neq 0\}$. Then,

$$T_x = \operatorname{span}\{(a_i)_{i \in \operatorname{supp}(x)}\}, \quad e_x[i] = \begin{cases} \operatorname{sign}(x[i]) & \text{if } i \in \operatorname{supp}(x) \\ 0 & \text{otherwise} \end{cases}, \;\text{ and }\; \sigma_{\mathcal{C}} = \|\cdot\|_\infty. \tag{6.2.7}$$

Thus if $\bar{x}$ is s-sparse, i.e. $|\operatorname{supp}(\bar{x})| = s$, then $\dim(T_{\bar{x}}) = s$ and $\|e_{\bar{x}}\|^2 = s$. Moreover

$$\mathbb{E}\left(\sigma_{\mathcal{C}}(g_{S_x})^2\right) = \mathbb{E}\left(\max_{i \in \operatorname{supp}(\bar{x})^c} |g[i]|^2\right),$$

which is the expectation of the maximum of $(n-s)$ χ^2-random variables with 1 degree of freedom. We then have, using [154, Lemma 3.2], that

$$\mathbb{E}\left(\max_{i \in \operatorname{supp}(\bar{x})^c} |g[i]|\right) \leq \left(\sqrt{2\log(n-s)} + 1\right)^2.$$

[4] This generalizes the reasoning of [154] beyond group sparsity.

Collecting the above in Lemma 6.2.12 and using Corollary 6.2.8, we obtain the following result.

Proposition 6.2.13. *Let $\bar{x}$ be an s-sparse vector. Let A be a matrix whose entries are i.i.d. $\mathcal{N}(0,1/m)$ with*

$$m \geq \frac{32(\nu+2)^2}{\nu^4} s \left(\left(\sqrt{2\log(n-s)} + 1 \right)^2 + 1 \right).$$

Then with probability at least $1 - 3e^{-\frac{\nu^2 m}{16}}$, the recovery of $\bar{x}$ (up to a global sign) is exact by solving $(\mathcal{P}_{\bar{y},0})$ with $R = \|\cdot\|_1$.

Remark 6.2.14. Clearly, $m \gtrsim s\log(n-s) + s$ measurements are sufficient for the exact recovery of an s-sparse vector from m phaseless measurements of a Gaussian map A. This can be improved to $m \gtrsim s\log(n/s) + s$ by exploiting the particular form of the normal cone of the ℓ_1 norm, see [61, Proposition 3.10]. This leads to a measurement bound similar to the one in [178]. Note however that their recovery guarantee is RIP-based, and thus is uniform over all s-sparse vectors while our recovery analysis is non-uniform.

$\ell_1 - \ell_2$ regularization The $\ell_1 - \ell_2$ regularization (a.k.a. group Lasso) is widely advocated to promote group/block sparsity, i.e. it drives all the coefficients in one group to zero together hence leading to group selection, see [47] for a comprehensive review. The group Lasso penalty with L groups reads

$$R(x) = \|x\|_{1,2} \stackrel{\text{def}}{=} \sum_{i=1}^{L} \|x[b_i]\|_2 . \tag{6.2.8}$$

where $\bigcup_{i=1}^{L} b_i = [\![n]\!]$, $b_i, b_j \subset [\![n]\!]$, and $b_i \cap b_j = \emptyset$ whenever $i \neq j$. Define the group support as $\mathrm{supp}_{\mathcal{B}}(x) \stackrel{\text{def}}{=} \{i \in [\![L]\!] : x[b_i] \neq 0\}$. Thus, one has

$$T_x = \mathrm{span}\{(a_j)_{\{j:\, \exists i \in \mathrm{supp}_{\mathcal{B}}(x), j \in b_i\}}\}, \quad e_x[b_i] = \begin{cases} \frac{x[b_i]}{\|x[b_i]\|_2} & \text{if } i \in \mathrm{supp}_{\mathcal{B}}(x) \\ 0 & \text{otherwise} \end{cases}, \tag{6.2.9}$$

and

$$\sigma_{\mathcal{C}}(v) = \max_{i \in \mathrm{supp}_{\mathcal{B}}(\bar{x})^c} \|v[b_i]\|_2 . \tag{6.2.10}$$

Thus if $\bar{x}$ is s-block sparse, i.e. $|\mathrm{supp}_{\mathcal{B}}(\bar{x})| = s$, and the groups have equal size B, we have $\dim(T_{\bar{x}}) = sB$ and $\|e_{\bar{x}}\|^2 = s$. Moreover, [154, Lemma 3.2] yields

$$\mathbb{E}\left(\max_{i \in \mathrm{supp}_{\mathcal{B}}(\bar{x})^c} \|g[b_i]\|^2 \right) \leq \left(\sqrt{2\log(L-s)} + \sqrt{B} \right)^2 .$$

Hence, we get the following result for the group Lasso.

Proposition 6.2.15. *Let $\bar{x}$ be an s-block sparse vector. Let A be a matrix whose entries are i.i.d. $\mathcal{N}(0,1/m)$ with*

$$m \geq \frac{32(\nu+2)^2}{\nu^4} s \left(\left(\sqrt{2\log(L-s)} + \sqrt{B} \right)^2 + B \right).$$

Then with probability at least $1 - 3e^{-\frac{\nu^2 m}{16}}$, the recovery of $\bar{x}$ (up to a global sign) is exact by solving $(\mathcal{P}_{\bar{y},0})$ with $R = \|\cdot\|_{1,2}$.

Remark 6.2.16. For group Lasso, in [101] the authors proposed the Copram algorithm which achieves reconstruction from $O(\frac{s^2}{B}\log(n))$. Our theoretical bound for recovery is of order $m \gtrsim s\,(2\log(L-s) + B)$. This bound is up to a constant similar to the case where the reconstruction is done from linear measurements and A is a Gaussian matrix [57, Theorem 3.1] and [153]. We think this is the first recovery bound for Group Lasso Phase retrieval. It shows the gap between the theoretical bound and the bound obtained using an iterative procedure to solve the group Lasso phase retrieval problem.

6.2.5 Recovery bounds for frame analysis-type regularizers

Analysis-type priors build upon the assumption that the signal of interest $\bar{x}$ is of low complexity (sparse) after being transformed by a so-called analysis operator. Given $D : \mathbb{R}^{n\times p}$, the analysis-type regularizer we consider is

$$R(x) = \gamma_{\mathcal{C}}(D^\top x). \tag{6.2.11}$$

where $\gamma_{\mathcal{C}}$ is a strong gauge (see (6.2.2) and the discussion just after). Since $\gamma_{\mathcal{C}}$ has a full domain, we have

$$\partial R(x) = D\partial\gamma_{\mathcal{C}}(D^\top x) = D\{v \in \mathbb{R}^p : \ v_{T_{D^\top x}} = e_{D^\top x} \ \text{ and } \ \sigma_{\mathcal{C}}(v_{S_{D^\top x}}) \leq 1\}, \tag{6.2.12}$$

where $e_{D^\top x}$ and $T_{D^\top x}$ are the model parameters of $\gamma_{\mathcal{C}}$ at $D^\top x$.

In this section, we will assume that D is a Parseval tight frame of $\mathbb{R}^n$, hence surjective, meaning that D is in the orthogonal group, i.e. $DD^\top = \mathrm{Id}_n$. Many popular sparsifying transforms in signal and image processing are Parseval tight frames (e.g. wavelets, curvelets, or concatenation of orthonormal bases; see [166]).

We can now state the following analysis-type prior version of Lemma 6.2.12.

Lemma 6.2.17. *Let R be of the form (6.2.2), where $\gamma_{\mathcal{C}}$ is a strong gauge and D is a Parseval tight frame. Let $z = D^\top g$ where $g \sim \mathcal{N}(0, \mathrm{Id}_n)$. Then for any $x \in \mathbb{R}^n \setminus \{0\}$*

$$w\left(\mathcal{D}_R(x) \cap \mathbb{S}^{n-1}\right) \leq \mathbb{E}\left(\sigma_{\mathcal{C}}(z_{S_{D^\top x}})\right) \|e_{D^\top x}\| + \sqrt{\dim\left(T_{D^\top x}\right)}. \tag{6.2.13}$$

The proof bears an apparent similarity with that of Lemma 6.2.12, but handling the presence of D necessitates new arguments.

Proof. Since D is surjective and $\gamma_{\mathcal{C}}$ is a strong gauge, we have $\mathrm{Argmin}(R) = \{0\}$. It then follows from [155, Theorem 23.7] that for any $x \neq 0$

$$\mathcal{N}_R(x) = \bigcup_{t\geq 0} t\partial R(x).$$

Combining this with (6.2.12), we get

$$w\left(\mathcal{D}_R(x) \cap \mathbb{S}^{n-1}\right) \leq \inf_{t\geq 0} \mathbb{E}\left(\mathrm{dist}\left(g, t\partial R(x)\right)\right) \leq \mathbb{E}\left(\mathrm{dist}\left(g, \tilde{t}D\partial\gamma_{\mathcal{C}}(D^\top x)\right)\right)$$

for any $\tilde{t} \geq 0$. Let us pick $v \in \mathbb{R}^p$ such that $v_{S_{D^\top x}} = z_{S_{D^\top x}}$ and $v_{T_{D^\top x}} = \sigma_{\mathcal{C}}(z_{S_{D^\top x}})e_{D^\top x}$. Obviously, $v \in \sigma_{\mathcal{C}}(z_{S_{D^\top x}})\partial\gamma_{\mathcal{C}}(D^\top x)$. However, although the entries of z are all standard Gaussian, they are not independent. We have

$$\begin{aligned}
w\left(\mathcal{D}_R(x) \cap \mathbb{S}^{n-1}\right) &\leq \mathbb{E}\left(\|g - Dv\|\right) \\
&= \mathbb{E}\left(\left\|DD^\top g - Dv\right\|\right) \\
&\leq \mathbb{E}\left(\|z - v\|\right) \\
&= \mathbb{E}\left(\left\|(z_{T_{D^\top x}} - v_{T_{D^\top x}}) + (z_{S_{D^\top x}} - v_{S_{D^\top x}})\right\|^2\right) \\
&= \mathbb{E}\left(\left\|z_{T_{D^\top x}} - \sigma_{\mathcal{C}}(z_{S_{D^\top x}})e_{D^\top x}\right\|\right) \\
&\leq \mathbb{E}\left(\sigma_{\mathcal{C}}(z_{S_{D^\top x}})\right) \|e_{D^\top x}\| + \mathbb{E}\left(\left\|z_{T_{D^\top x}}\right\|\right).
\end{aligned}$$

In the first equality we used that D is a Parseval tight frame, and in the second inequality that

$\|D\| \leq 1$. Let $s(M) \in \mathbb{R}^p$ be the decreasing sequence of singular values of M. We have

$$\begin{aligned}
\mathbb{E}\left(\left\|z_{T_{D^\top x}}\right\|\right)^2 &\leq \mathbb{E}\left(\left\|z_{T_{D^\top x}}\right\|^2\right) \\
&= \mathbb{E}\left(\operatorname{tr}\left(\mathrm{P}_{T_{D^\top x}} D^\top g g^\top D \mathrm{P}_{T_{D^\top x}}\right)\right) \\
&= \operatorname{tr}\left(\mathrm{P}_{T_{D^\top x}} D^\top D \mathrm{P}_{T_{D^\top x}}\right) \\
&= \operatorname{tr}\left(\mathrm{P}_{T_{D^\top x}} D^\top D\right) \\
&\leq \left\langle s\left(\mathrm{P}_{T_{D^\top x}}\right), s(D^\top D)\right\rangle \\
&\leq \left\|s\left(\mathrm{P}_{T_{D^\top x}}\right)\right\|_1 \|D\|^2 \\
&= \dim\left(T_{D^\top x}\right).
\end{aligned}$$

In the first inequality, we used Jensen's inequality. In the second one, we invoked a well-known result essentially due to von Neumann [177]. In the third one, we used Hölder's inequality, and the last equality uses that D is a Parseval tight frame and standard properties of orthogonal projectors on subspaces. □

The challenging part to compute the upper-bound in (6.2.13) is to compute the expectation therein while the entries of z which are not independent (except the obvious case where D is orthonormal). For the case of the ℓ_1 norm, however, this can be achieved.

ℓ_1 frame analysis regularization In this case, $\gamma_{\mathcal{C}} = \|\cdot\|_1$, and thus

$$T_{D^\top x} = \operatorname{span}\{(a_i)_{i \in \operatorname{supp}(D^\top x)}\}, \quad e_{D^\top x}[i] = \begin{cases} \operatorname{sign}((D^\top x)[i]) & \text{if } i \in \operatorname{supp}(D^\top x) \\ 0 & \text{otherwise} \end{cases}, \text{ and } \sigma_{\mathcal{C}} = \|\cdot\|_\infty. \tag{6.2.14}$$

Thus if $\bar{x}$ is s-sparse in the dictionary $D^\top$, i.e. $\left|\operatorname{supp}(D^\top \bar{x})\right| = s$, then $\dim(T_{D^\top \bar{x}}) = s$ and $\|e_{D^\top \bar{x}}\|^2 = s$. Moreover

$$\mathbb{E}\left(\sigma_{\mathcal{C}}(z_{S_x})\right) = \mathbb{E}\left(\max_{i \in \operatorname{supp}(D^\top \bar{x})^c} |z[i]|\right).$$

A standard estimate of the expectation of the ℓ_∞ norm of (not necessarily independent) standard Gaussian random vectors gives

$$\mathbb{E}\left(\max_{i \in \operatorname{supp}(D^\top \bar{x})^c} |z[i]|\right) \leq \sqrt{2\log(2(p-s))}.$$

Inserting the above in Lemma 6.2.17, and using Corollary 6.2.8 together with Jensen's inequality, we get the following.

Proposition 6.2.18. *Let $\bar{x}$ such that $D^\top \bar{x}$ is s-sparse. Let A be a matrix whose entries are i.i.d. $\mathcal{N}(0, 1/m)$ with*

$$m \geq \frac{64(\nu+2)^2}{\nu^4} s\left(2\log(2(p-s)) + 1\right).$$

Then with probability at least $1 - 3e^{-\frac{\nu^2 m}{16}}$, the recovery of $\bar{x}$ (up to a global sign) is exact by solving ($\mathcal{P}_{\bar{y},0}$) with $R = \left\|D^\top \cdot\right\|_1$.

Remark 6.2.19.

- Consequently, it is sufficient to have $m \gtrsim s\log(n-s) + s$ to ensure the exact recovery of a vector, which is s-sparse in a tight frame, from m phaseless measurements of a Gaussian map A. We are not aware of any such result in the phase recovery literature. Observe also that the sample complexity bound we get is nearly (up to constants) the same as for exact recovery from linear Gaussian measurements [57].

- Unlike for the ℓ_1 case, when the gauge is the Group Lasso it is quite challenging to compute the Gaussian width of the descent cone as the entries of $z = D^\top g$ are dependents. The difficulty lies in bounding the term $\mathbb{E}\left(\sigma_{\mathcal{C}}(z_{S_{D^\top x}})\right)$ which, in this case, contains the square of dependant chi-variable over the maximal sub-blocks and we don't know how to concentrate them.

6.2.6 Recovery bounds for total variation

Total variation (TV) corresponds to the case where the analysis operator $D^\top$ in (6.2.11) is the (discrete) gradient ∇ and $\gamma_{\mathcal{C}} = \|\cdot\|_1$. In the 1D case, TV regularization reads

$$R(x) = \|\nabla x\|_1, \quad \text{where} \quad \nabla x[i] = x[i+1] - x[i], \quad \text{for} \quad i = 1, 2, \ldots, n-1.$$

R promotes signals x whose gradient is sparse, $|\mathrm{supp}(\nabla x)| \leq s$, or in other words, signals that are piecewise constant with at most s jumps.

Bounding the Gaussian width of the descent cone of R in this case is very challenging as ∇ has a non-trivial kernel, and thus does not fit within the setting of the previous section. However, if the jumps of an s-gradient sparse signal x are well separated, [81] proposed a non-trivial construction of the dual vector to compute the Gaussian width of the descent cone of the Total Variation. More precisely, assume that there exists $\Delta > 0$ such that

$$\min_{i \in [\![s+1]\!]} \frac{|k_i - k_{i-1}]}{n} \geq \frac{\Delta}{s+1},$$

where $\mathrm{supp}(\nabla x) = \{k_1, \ldots, k_s\}$ with $0 = k_0 < k_1 < \ldots < k_s < k_{s+1} = n$. It was shown in [81, Theorem 2.10] that if $\Delta \geq 8s/n$, then

$$w\left(\mathcal{D}_{\|\nabla \cdot\|_1}(x) \cap \mathbb{S}^{n-1}\right)^2 \leq \frac{C}{\Delta} s \log(n)^2,$$

for some numerical constant $C > 0$.

We are then able to state the following result.

Proposition 6.2.20. *Let $\bar{x}$ be an s-group sparse vector such that its separation constant verifies $\Delta \geq 8s/n$. Let A be a matrix whose entries are i.i.d. $\mathcal{N}(0, 1/m)$ with*

$$m \gtrsim \frac{1}{\Delta} s \log(n)^2.$$

Then with probability at least $1 - 3e^{-\frac{\nu^2 m}{16}}$, the recovery of $\bar{x}$ (up to a global sign) is exact by solving ($\mathcal{P}_{\bar{y},0}$) with $R = \|\nabla \cdot\|_1$.

Remark 6.2.21. As mentioned before, finding complexity bounds for TV minimization is quite challenging even in the compressed sensing literature. In this setting, [138, 139, 49] showed, for two or higher dimensions signals, robust and stable recovery when A is Gaussian and composed with orthonormal Haar wavelet transform. The complexity in this case is of order $m \geq s\mathrm{PolyLog}(n, s)$. The success of this approach relies on establishing a connection between the compressibility of Haar wavelet representations and the bounded variation of a function and this does not hold in one dimension. We think it is possible to extend this result to the case of phase retrieval and we leave this as future work.

It was shown in [49] that for general one-dimension signals, it is not possible to recover the signal from $m \leq C_1\sqrt{sn} - C_2$. This is why the authors of [81] consider signals where the jumps are more separated and thus achieving $m \geq s\log(n)^2$ which is of interest in this work. It would be also interesting to explore the case of Fourier sub-sampled measurements for phaseless measurements. Indeed for compress sensing, [103, 152] use variable-density sampling of the Fourier transform, wherein sampling in the low frequencies is denser than in the high frequencies to provide complexity bound. We wind up this discussion by noticing that this result is new for phase retrieval.

6.3 Stable Recovery: Constrained Problem

When we have access only to inaccurate noisy measurements as in (NoisyPR), a natural formulation is one in which the equality constraint in ($\mathcal{P}_{\bar{y},0}$) is relaxed to an inequality constraint leading to

$$\inf_{x\in\mathbb{R}^n} R(x) \quad \text{s.t.} \quad \left\| y - |Ax|^2 \right\| \leq \rho, \tag{$\mathcal{P}_{y,\rho}$}$$

where ρ is an upper bound on the size of the noise ϵ. In the inverse problems literature, this formulation is known as the residual method or Mozorov regularization. In the following, we denote $\overline{\mathcal{F}}_{y,\rho} \overset{\text{def}}{=} \{w \in \mathbb{R}^n : \ \|y - |w|^2\| \leq \rho\}$. We obviously have $\overline{\mathcal{F}}_{\bar{y},0} = \overline{\mathcal{F}}$. We also use the shorthand notation $\mathcal{S}_{y,\rho}$ for the set of minimizers of ($\mathcal{P}_{y,\rho}$).

We start by showing that ($\mathcal{P}_{y,\rho}$) has minimizers. This result does not require convexity of R.

Proposition 6.3.1. *Let $R : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper and lsc function. Assume that $A(\mathrm{dom}(R)) \cap \overline{\mathcal{F}}_{y,\rho} \neq \emptyset$, and that assumptions (ii)-(iii) of Proposition 6.2.1 hold. Then problem ($\mathcal{P}_{y,\rho}$) has a non-empty compact set of minimizers.*

Proof. The proof is similar to that of Proposition 6.2.1 replacing $\overline{\mathcal{F}}$ by $\overline{\mathcal{F}}_{y,\rho}$, and the latter is a compact set. □

We are now ready to state our (deterministic) stability result.

Theorem 6.3.2. *Consider the noisy phaseless measurements in (NoisyPR) where $\|\epsilon\| \leq \rho$. Assume that R verifies (**H.1**). Then, for any $x^\star_{y,\rho} \in \mathcal{S}_{y,\rho}$, we have*

$$\mathrm{dist}(x^\star_{y,\rho}, \overline{\mathcal{X}}) \leq \frac{2\rho}{s_{\min}},$$

where

$$s_{\min} \overset{\text{def}}{=} \inf \left\{ \min_{I\subset[\![m]\!], |I|\geq m/2} \left\| A^I z \right\| : \ z \in \mathcal{D}_R(\bar{x}) \cap \mathbb{S}^{n-1} \right\}.$$

Proof. The proof as a flavour of the reasoning in the proof of Theorem 6.2.4. Let $I \subset [\![m]\!]$ such that $\left\langle a_r, x^\star_{y,\rho} \right\rangle = \langle a_r, \bar{x}\rangle$ for all $r \in I$, and I^c its complement where the inner products have opposite signs. Thus either $|I| \geq m/2$ or $|I^c| \geq m/2$. Assume that $|I| \geq m/2$. Then

$$\begin{aligned}\left\| |Ax^\star_{y,\rho}| - |A\bar{x}| \right\|^2 &= \left\| |A^I x^\star_{y,\rho}| - |A^I \bar{x}| \right\|^2 + \left\| |A^{I^c} x^\star_{y,\rho}| - |A^{I^c} \bar{x}| \right\|^2 \\ &\geq \left\| A^I x^\star_{y,\rho} - A^I \bar{x} \right\|^2.\end{aligned}$$

Recall that $\bar{x} \in \overline{\mathcal{F}}_{y,\rho}$ by assumption on the noise. Thus $R(x^\star_{y,\rho}) \leq R(\bar{x})$ and in turn $x^\star_{y,\rho} - \bar{x} \in \mathcal{D}_R(\bar{x})$. Therefore,

$$\left\| |Ax^\star_{y,\rho}| - |A\bar{x}| \right\|^2 \geq s^2_{\min} \left\| x^\star_{y,\rho} - \bar{x} \right\|^2.$$

For the case where $|I^c| \geq m/2$, we argue similarly to infer that

$$\left\| |Ax^\star_{y,\rho}| - |A\bar{x}| \right\|^2 \geq s^2_{\min} \left\| x^\star_{y,\rho} + \bar{x} \right\|^2.$$

Overall, we have

$$\mathrm{dist}(x^\star_{y,\rho}, \overline{\mathcal{X}}) \leq \frac{\left\| |Ax^\star_{y,\rho}| - |A\bar{x}| \right\|}{s_{\min}} \leq \frac{\left\| y - |Ax^\star_{y,\rho}| \right\| + \|\epsilon\|}{s_{\min}} \leq \frac{2\rho}{s_{\min}}.$$

□

When A is a standard Gaussian map, we obtain the following general error bound.

Proposition 6.3.3. *Consider the noisy phaseless measurements in* (NoisyPR) *where* $\|\epsilon\| \leq \rho$. *Suppose that* (**H.1**) *holds. Let* ν *be as defined in Lemma* 6.2.6 *and* A *be a matrix whose entries are i.i.d.* $\mathcal{N}(0, 1/m)$ *with*

$$m \geq \frac{32(\nu+2)^2}{\nu^4} w\left(\mathcal{D}_R(\bar{x}) \cap \mathbb{S}^{n-1}\right)^2.$$

Then with probability at least $1 - 3e^{-\frac{\nu^2 m}{16}}$, *the following statement holds: for any* $x^\star_{y,\rho} \in \mathcal{S}_{y,\rho}$,

$$\operatorname{dist}(x^\star_{y,\rho}, \overline{\mathcal{X}}) \leq \frac{4\rho}{\nu(1-1/\sqrt{2})}.$$

Proof. From the proof of Corollary 6.2.8, we have

$$s_{\min} \geq \nu/2(1-1/\sqrt{2}),$$

with probability at least $1 - 3e^{-\frac{\nu^2 m}{16}}$ under the bound on m. Combining this with Theorem 6.2.4, we conclude. □

Remark 6.3.4.

- In [80], the authors studied the stability of ℓ_1-norm phase retrieval against noise they showed that for $m \geq s\log(n/s)$ we can stably reconstruct a $s-$sparse signal for measurements that satisfies the strong-RIP property. Our stability result here goes far beyond the ℓ_1-norm and does not require that A satisfies the strong-RIP but rather a structural assumption. For Gaussian measurements, Proposition 6.3.3 entails the recovery bound depending on the descent cone.
- We can easily instantiate the last result for the regularizers studied in Section 6.2.4, 6.2.5 and 6.2.6, which in turn will give sample complexity bounds for the error bound of Theorem 6.2.7 to hold.
- Let us notice that despite the nice stability properties of ($\mathcal{P}_{y,\rho}$), it is not clear if it can be solved with an efficient algorithmic scheme. Indeed, although R is convex, the constraint in ($\mathcal{P}_{y,\rho}$) is highly non-convex, and it is very challenging to project onto it. On the other hand, as stated in the introduction, ($\mathcal{P}_{y,\lambda}$) is amenable to the efficient Bregman proximal gradient algorithmic scheme proposed [40] and further studied in Chapter 5. This is the reason we now turn our attention to ($\mathcal{P}_{y,\lambda}$).

6.4 Stable Recovery: Penalized Problem

We now turn to study the noise-aware problem ($\mathcal{P}_{y,\lambda}$). In particular, the following questions will be of most interest to us:

- Convergence: this ensures that for $\lambda \to 0$ as $\epsilon \to 0$, the set of regularized solutions converges to either $\bar{x}$ or $-\bar{x}$.
- Convergence rates: this provides an estimate of the rate at which the above convergence takes place.

As will see, studying the stability of ($\mathcal{P}_{y,\lambda}$) is more involved than for ($\mathcal{P}_{y,\rho}$). One of the main difficulties, which was also highlighted for linear inverse problems (see [174]), is that a minimizer of ($\mathcal{P}_{y,\lambda}$) is not anymore in the descent cone of R at $\bar{x}$.

In this section, we set

$$\mathcal{S}_{y,\lambda} \overset{\text{def}}{=} \operatorname*{Argmin}_{x \in \mathbb{R}^n} F_{y,\lambda}(x),$$

where we recall the objective $F_{y,\lambda}$ from ($\mathcal{P}_{y,\lambda}$).

We begin by providing conditions for the existence of minimizers. Again, this does not need convexity of R.

Proposition 6.4.1. *Let $R : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper and lsc function. Assume that assumptions (ii)-(iii) of Proposition 6.2.1 hold. Then for any $\lambda > 0$ and $y \in \mathbb{R}^m$, problem $(\mathcal{P}_{y,\lambda})$ has a non-empty compact set of minimizers.*

Proof. The proof is similar to that of Proposition 6.2.1 replacing $\iota_{\overline{\mathcal{F}}}$ by $\|y - |\cdot|^2\|^2/2$, and the latter turns out to be a smooth and coercive function. We omit the details for the sake of brevity. □

6.4.1 Convergence

We start by proving the following convergence result for any minimizer $x^\star_{y,\lambda}$ of $(\mathcal{P}_{y,\lambda})$.

Theorem 6.4.2. *Consider the noisy phaseless measurements in (NoisyPR). Let $\sigma \overset{\text{def}}{=} \|\epsilon\|$. Assume that (H.1), (H.2) and assumptions (ii)-(iii) of Proposition 6.2.1 hold. Suppose also that*

$$\lambda \to 0 \ \text{ and } \ \sigma^2/\lambda \to 0, \quad \text{as} \quad \sigma \to 0.$$

Then,

$$|Ax^\star_{y,\lambda}| \to |A\bar{x}|, \quad R(x^\star_{y,\lambda}) \to R(\bar{x}) \ \text{ and } \ \operatorname{dist}\left(x^\star_{y,\lambda}, \overline{\mathcal{X}}\right) \to 0 \quad \text{as} \quad \sigma \to 0.$$

Proof. Let $y_k = \bar{y} + \epsilon_k$, $\sigma_k = \|\epsilon_k\|$ with $\sigma_k \to 0$ as $k \to +\infty$. Observe that for any y_k and $\lambda_k > 0$ $\mathcal{S}_{y_k,\lambda_k}$ is a non-empty compact set thanks to Proposition 6.4.1. Let $x^\star_k \in \mathcal{S}_{y_k,\lambda_k}$. We have by optimality that

$$\left\|y_k - |Ax^\star_k|^2\right\|^2 + \lambda_k R(x^\star_k) \leq \|y_k - \bar{y}\|^2 + \lambda_k R(\bar{x}) = \sigma_k^2 + \lambda_k R(\bar{x}).$$

Thus

$$\left\|y_k - |Ax^\star_k|^2\right\|^2 \leq \lambda_k \left(\sigma_k^2/\lambda_k + R(\bar{x})\right)$$

and

$$R(x^\star_{y_k,\lambda_k}) \leq \sigma_k^2/\lambda_k + R(\bar{x}).$$

In turn,

$$\left\||Ax^\star_k|^2 - \bar{y}\right\|^2 \leq 2\left(\left\|y_k - |Ax^\star_k|^2\right\|^2 + \sigma_k^2\right) \leq 2\left(\lambda_k\left(\sigma_k^2/\lambda_k + R(\bar{x})\right) + \sigma_k^2\right).$$

Since the right hand side of this inequality goes to 0 as $k \to +\infty$, we deduce that

$$\lim_{k\to+\infty} |Ax^\star_k|^2 = \bar{y}. \tag{6.4.1}$$

Moreover,

$$\limsup_{k\to+\infty} R(x^\star_k) \leq R(\bar{x}). \tag{6.4.2}$$

We therefore obtain

$$\limsup_{k\to+\infty} F_{\bar{y},1}(x^\star_k) = \limsup_{k\to+\infty}\left(\left\||Ax^\star_k|^2 - \bar{y}\right\|^2 + R(x^\star_k)\right) \leq R(\bar{x})$$

This means that there exists $k_0 \in \mathbb{N}$ such that $(x^\star_k)_{k\geq k_0}$ belongs to the sublevel set of $F_{\bar{y},1}$ at $2R(\bar{x})$, that we denote $\mathcal{C}_F$. Since $F_{\bar{y},1}$ is lsc and coercive under our assumptions, its sublevel sets are compact and so is $\mathcal{C}_F$. In turn, $(x^\star_k)_{k\geq k_0}$ lives on the compact set $\mathcal{C}_F$. The sequence thus possesses a convergent subsequence and every accumulation point lies $\mathcal{C}_F$. Let $(x^\star_{k_j})_{j\in\mathbb{N}}$ be a convergent subsequence, say $x^\star_{k_j} \to x^*$. We have $|Ax^\star_{k_j}|^2 \to |Ax^*|^2$, and in view of (6.4.1), we obtain

$$|Ax^*|^2 = \bar{y}.$$

Moreover, by lower-semicontinuity and (6.4.2),

$$R(x^*) \leq \liminf_{j\to+\infty} R(x_{k_j}) \leq \limsup_{j\to+\infty} R(x^\star_{k_j}) \leq R(\bar{x}). \tag{6.4.3}$$

Invoking Theorem 6.2.4 (which holds under our assumptions), the last two relations mean that $x^* \in \mathcal{S}_{\bar{y},0} = \overline{\mathcal{X}}$. The latter relation also shows that $R(\bar{x}) = R(x^*)$, and thus we have $R(x_{k_j}) \to R(\bar{x})$. Since this holds for any convergent subsequence, we conclude. □

For Gaussian measurements, combining Theorem 6.4.2 and Corollary 6.2.8, we have the following asymptotic robustness result provided m is large enough.

Proposition 6.4.3. *Consider the noisy phaseless measurements in (NoisyPR) and let* $\sigma \overset{\text{def}}{=} \|\epsilon\|$. *Assume that* **(H.1)** *and assumptions (ii)-(iii) of Proposition 6.2.1 hold. Suppose also that*

$$\lambda \to 0 \text{ and } \sigma^2/\lambda \to 0, \quad \text{as} \quad \sigma \to 0.$$

Let ν *be as defined in Lemma 6.2.6, and* A *be a matrix whose entries are i.i.d.* $\mathcal{N}(0, 1/m)$ *with*

$$m \geq \frac{32(\nu+2)^2}{\nu^4} w\left(\mathcal{D}_R(\bar{x}) \cap \mathbb{S}^{n-1}\right)^2.$$

Then with probability at least $1 - 3e^{-\frac{\nu^2 m}{16}}$,

$$\operatorname{dist}\left(x^\star_{y,\lambda}, \overline{\mathcal{X}}\right) \to 0 \quad \text{as} \quad \sigma \to 0.$$

This result can be specialized with the corresponding sample complexity bounds for each of the regularizers considered in Section 6.2.4 to 6.2.6. We leave the details to the reader.

6.4.2 Deterministic convergence rate

We now turn to quantifying the rate at which convergence of Theorem 6.4.2 occurs. This will be possible under more stringent conditions. For instance, we will require the noise to be small enough to that the rate is actually local. We moreover need a non-degeneracy condition and a restricted injectivity conditions which are standard in inverse problems; see Remark 6.4.5 for a detailed discussion.

To lighten notation, let us denote

$$B_{\bar{x}} \overset{\text{def}}{=} \operatorname{diag} A\bar{x} A.$$

Although the following result can be stated for general symmetric convex regularizers R, to avoid additional technicalities and make the presentation simpler, we will restrict our attention to the case of analysis-type symmetric strong gauges which will be sufficient for our purposes. More precisely, R will be of the form (6.2.11), where D is a Parseval tight frame and $\gamma_{\mathcal{C}}$ is a symmetric strong gauge. We recall the definition, notations, and properties of Section 6.2.4.

Theorem 6.4.4. *Consider the noisy phaseless measurements in (NoisyPR). Let* $\sigma \overset{\text{def}}{=} \|\epsilon\|$. *Assume that* R *is as in (6.2.11), where* D *is a Parseval tight frame and* $\gamma_{\mathcal{C}}$ *is a symmetric strong gauge, and that*

$$\exists q \in \mathbb{R}^m \quad s.t. \quad {B_{\bar{x}}}^\top q \in \operatorname{ri}(\partial R(\bar{x})) \tag{6.4.4}$$

and

$$\ker(B_{\bar{x}}) \cap \operatorname{Im}(D_{T_{D^\top \bar{x}}}) = \{0\}. \tag{6.4.5}$$

Consider the choice $\lambda = c\sigma$, *for some* $c > 0$. *Then, for* σ *small enough and any minimizer* $x^\star_{y,\lambda} \in \mathcal{S}_{y,\lambda}$, *we have*

$$\operatorname{dist}\left(x^\star_{y,\lambda}, \overline{\mathcal{X}}\right) \leq C\sigma,$$

where $C > 0$ *is a constant which depends in particular on* A, $T_{D^\top \bar{x}}$, c *and* q.

A few remarks are in order before we proceed with the proof.

Remark 6.4.5.

- The error bound of Theorem 6.4.4 tells us that for small noise, the distance of any minimizer of $(\mathcal{P}_{y,\lambda})$ to $\overline{\mathcal{X}}$ is within a factor of the noise level, which justifies the terminology "linear convergence rate" known in the inverse problem literature. On the other hand, since R verifies all assumptions of Theorem 6.4.2, we have that $\mathcal{S}_{y,\lambda}$ are bounded uniformly in (y,λ), and it follows from (6.4.6) that the error $\left\| |Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2 \right\|$ is global and scales as $O(\max(\sigma^{1/2},\sigma))$.
- Source condition: condition (6.4.4) is a strengthened version of the so-called "source condition" or "range condition" in the literature of inverse problems; see [158] for a general overview of this condition and its implications. In this case, $v = B^\top_{\bar{x}} q$ is called a non-degenerate dual certificate[5]; see [174] for a detailed discussion in the case of linear inverse problems.
- Restricted injectivity: the restricted injectivity condition (6.4.5) is only favorable when $\gamma_{\mathcal{C}}$ is non-smooth at $D^\top \bar{x}$, hence the intuition that $\gamma_{\mathcal{C}}$ (hence R) promotes low-dimensional vectors. Indeed, the higher the degree of non-smoothness, the lower the dimension of the subspace $T_{D^\top \bar{x}}$, and hence the less number of measurements is needed for the restricted injectivity to hold. From the calculus rules in [173, Proposition 10(i)-(ii)], the model subspace of the regularizer R at $\bar{x}$ is $\ker(D^\top_{S_{D^\top \bar{x}}})$. Since D is a Parseval tight frame, one can easily show that $\ker(D^\top_{S_{D^\top \bar{x}}}) \subseteq \mathrm{Im}(D_{T_{D^\top \bar{x}}})$, with equality if D orthonormal. Thus (6.4.5) implies that $B_{\bar{x}}$ is injective on the model subspace of R at $\bar{x}$, which is a minimal requirement to ensure recovery as is known even for linear inverse problems.
- $B_{\bar{x}}$ is nothing but the Jacobian of the non-linear mapping $x \in \mathbb{R}^n \mapsto |Ax|^2/2$. Convergence rates for regularized non-linear inverse problems were studied in [158]. However, their conditions are too stringent and do not hold for the case of phase retrieval by solving $(\mathcal{P}_{y,\lambda})$.

The following lemma is a key step towards establishing our error bound.

Lemma 6.4.6. *Let R as in (6.2.11) where $D \in \mathbb{R}^{p\times n}$ and $\gamma_{\mathcal{C}}$ is a strong gauge of $\mathcal{C}$. Let $x \in \mathbb{R}^n$. Then, for any $w \in \mathrm{ri}(\gamma_{\mathcal{C}}(D^\top x))$ and $z \in \mathbb{R}^n$*

$$\gamma_{\mathcal{C}}\left(\mathrm{P}_{S_{D^\top x}} D^\top (z-x)\right) \leq \frac{D^{Dw}_R(z,x)}{1-\sigma_{\mathcal{C}}\left(w_{S_{D^\top x}}\right)} = \frac{D^w_{\gamma_{\mathcal{C}}}(D^\top z, D^\top x)}{1-\sigma_{\mathcal{C}}\left(w_{S_{D^\top x}}\right)}.$$

Observe that by the decomposability property in (6.2.3), $w \in \mathrm{ri}(\partial\gamma_{\mathcal{C}}(D^\top x))$ is equivalent to

$$w_{T_{D^\top z}} = e_{D^\top x} \ \text{ and } \ \sigma_{\mathcal{C}}(w_{S_{D^\top x}}) < 1.$$

In plain words, the denominator in Lemma 6.4.6 does not vanish and the statement is not vacuous. In fact, this denominator can be viewed as a "distance" to degeneracy.

Proof. We start by noting that $\mathrm{ri}(\partial R(x)) = D\mathrm{ri}(\partial\gamma_{\mathcal{C}}(D^\top x))$. Let $v = Dw$. We have by convexity and decomposability of the subdifferential of $\gamma_{\mathcal{C}}$ that for any pair $(u,w) \in \partial\gamma_{\mathcal{C}}(D^\top x) \times \mathrm{ri}\left(\partial\gamma_{\mathcal{C}}(D^\top x)\right)$,

$$\begin{aligned} D^v_R(z,x) &= D^w_{\gamma_{\mathcal{C}}}(D^\top z, D^\top x) \\ &\geq D^w_{\gamma_{\mathcal{C}}}(D^\top z, D^\top x) - D^u_{\gamma_{\mathcal{C}}}(D^\top z, D^\top x) \\ &= \left\langle u-w, D^\top z - D^\top x \right\rangle \\ &= \left\langle u_{S_{D^\top z}} - w_{S_{D^\top z}}, D^\top z - D^\top x \right\rangle. \end{aligned}$$

From [130, Theorem 1], specialized to strong gauges, we have that for any $\omega \in \mathbb{R}^p$, $\exists \tilde{u} \in \partial\gamma_{\mathcal{C}}(D^\top x)$ such that

$$\gamma_{\mathcal{C}}(\omega_{S_{D^\top x}}) = \left\langle \tilde{u}_{S_{D^\top x}}, \omega_{S_{D^\top x}} \right\rangle.$$

[5] Strictly speaking, the terminology "dual" may seem awkward because of non-convexity of the phase retrieval problem $(\mathcal{P}_{y,\lambda})$ though it is weakly convex.

Applying this with $\omega = D^\top z - D^\top x$ and taking $u = \tilde{u}$, continuing the above chain of inequalities yields

$$\begin{aligned} D_R^v(z,x) &\geq \gamma_{\mathcal{C}}(\mathrm{P}_{S_{D^\top x}}(D^\top z - D^\top x)) - \left\langle w_{S_{D^\top x}}, \mathrm{P}_{S_{D^\top x}}(D^\top z - D^\top x)\right\rangle \\ &\geq \gamma_{\mathcal{C}}(\mathrm{P}_{S_{D^\top x}}(D^\top z - D^\top x))\left(1 - \sigma_{\mathcal{C}}(w_{S_{D^\top x}})\right), \end{aligned}$$

where in the last inequality, we used the duality inequality which holds by polarity between $\gamma_{\mathcal{C}}$ and $\sigma_{\mathcal{C}}$. This concludes the proof. □

Proof of Theorem 6.4.4. By symmetry of R, it can be easily seen that $\partial R(-\bar{x}) = -\partial R(\bar{x})$, and thus $T_{-D^\top \bar{x}} = T_{D^\top \bar{x}}$ and $e_{-D^\top \bar{x}} = -e_{D^\top \bar{x}}$. Therefore, if (6.4.4)-(6.4.5) hold then so they do at $-\bar{x}$ and vice versa.

Let $x^\star_{y,\lambda} \in \mathcal{S}_{y,\lambda}$ and suppose that $\bar{x}$ is its closest point in $\overline{\mathcal{X}}$. We have by optimality that

$$\left\|y - |Ax^\star_{y,\lambda}|^2\right\|^2 + \lambda R(x^\star_{y,\lambda}) \leq \sigma^2 + \lambda R(\bar{x}).$$

By the source condition (6.4.4), there exists $q \in \mathbb{R}^m$ such that $v = Dw \overset{\text{def}}{=} B_{\bar{x}}^\top q \in \mathrm{ri}(\partial R(\bar{x}))$. In turn, $w \in \mathrm{ri}(\partial \gamma_{\mathcal{C}}(D^\top \bar{x}))$. Convexity of R then implies

$$\begin{aligned} \left\|y - |Ax^\star_{y,\lambda}|^2\right\|^2 + \lambda D_R^v(x^\star_{y,\lambda}, \bar{x}) &\leq \sigma^2 - \lambda \left\langle q, B_{\bar{x}}(x^\star_{y,\lambda} - \bar{x})\right\rangle \\ &= \sigma^2 + \frac{\lambda}{2}\left\langle q, |Ax^\star_{y,\lambda} - A\bar{x}|^2 + (|Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2)\right\rangle \\ &\leq \sigma^2 + \frac{\lambda}{2}\left(\|q\|\left\|Ax^\star_{y,\lambda} - A\bar{x}\right\|_4^2 + \|q\|\left\||Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2\right\|\right) \\ &\leq \sigma^2 + \frac{\lambda}{2}\|q\|\left(\left\|Ax^\star_{y,\lambda} - A\bar{x}\right\|^2 + \left\||Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2\right\|\right). \end{aligned}$$

Strong convexity of $\|\cdot\|^2$ implies that

$$\left\|y - |Ax^\star_{y,\lambda}|^2\right\|^2 - \sigma^2 \geq 2\left\langle \epsilon, |Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2\right\rangle + \left\||Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2\right\|^2$$

Thus

$$\begin{aligned} &\left\||Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2\right\|^2 + \lambda D_R^v(x^\star_{y,\lambda}, \bar{x}) \\ &\leq -2\left\langle \epsilon, |Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2\right\rangle + \frac{\lambda}{2}\|q\|\left(\left\|Ax^\star_{y,\lambda} - A\bar{x}\right\|^2 + \left\||Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2\right\|\right) \\ &\leq \frac{\lambda}{2}\|q\|\left\|Ax^\star_{y,\lambda} - A\bar{x}\right\|^2 + \left(2\sigma + \frac{\lambda}{2}\|q\|\right)\left\||Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2\right\| \\ &\leq \frac{\lambda}{2}\|q\|\|A\|^2\left\|x^\star_{y,\lambda} - \bar{x}\right\|^2 + \frac{\left(2\sigma + \frac{\lambda}{2}\|q\|\right)^2}{2} + \frac{1}{2}\left\||Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2\right\|^2, \end{aligned}$$

and therefore

$$\left\||Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2\right\|^2 + 2\lambda D_R^v(x^\star_{y,\lambda}, \bar{x}) \leq \lambda\|q\|\|A\|^2\left\|x^\star_{y,\lambda} - \bar{x}\right\|^2 + \left(2\sigma + \frac{\lambda}{2}\|q\|\right)^2.$$

With the choice of λ and non-negativity of the Bregman divergence of R, we get

$$\left\||Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2\right\| \leq \left(c^{1/2}\|q\|^{1/2}\|A\|\left\|x^\star_{y,\lambda} - \bar{x}\right\|\right)\sigma^{1/2} + \left(2 + \frac{c}{2}\|q\|\right)\sigma$$

and (6.4.6)

$$D_R^v(x^\star_{y,\lambda}, \bar{x}) \leq \frac{\|q\|}{2}\|A\|^2\left\|x^\star_{y,\lambda} - \bar{x}\right\|^2 + \frac{(2 + \frac{c}{2}\|q\|)^2}{2c}\sigma.$$

By the triangle inequality and since D is a Parseval tight frame, we get

$$\begin{aligned} \left\|x^\star_{y,\lambda} - \bar{x}\right\| &= \left\|DD^\top(x^\star_{y,\lambda} - \bar{x})\right\| \\ &\leq \left\|D\mathrm{P}_{T_{D^\top \bar{x}}}D^\top(x^\star_{y,\lambda} - \bar{x})\right\| + \left\|\mathrm{P}_{S_{D^\top \bar{x}}}D^\top(x^\star_{y,\lambda} - \bar{x})\right\|. \end{aligned}$$

Denote $V_{\bar{x}} \overset{\text{def}}{=} \mathrm{Im}(D_{T_{D^\top \bar{x}}})$. In view of (6.4.5), we have $B^+_{\bar{x}V_{\bar{x}}} = (B^\top_{\bar{x}V_{\bar{x}}} B_{\bar{x}V_{\bar{x}}})^{-1} B^\top_{\bar{x}V_{\bar{x}}}$. Noting that $D\mathrm{P}_{T_{D^\top \bar{x}}} D^\top (x^\star_{y,\lambda} - \bar{x}) \in V_{\bar{x}}$ and using Lemma 6.4.6, we obtain

$$
\begin{aligned}
&\left\|x^\star_{y,\lambda} - \bar{x}\right\| \\
&\leq \left\|B^+_{\bar{x}V_{\bar{x}}} B_{\bar{x}V_{\bar{x}}} D\mathrm{P}_{T_{D^\top \bar{x}}} D^\top (x^\star_{y,\lambda} - \bar{x})\right\| + \left\|\mathrm{P}_{S_{D^\top \bar{x}}} D^\top (x^\star_{y,\lambda} - \bar{x})\right\| \\
&= \left\|B^+_{\bar{x}V_{\bar{x}}} B_{\bar{x}} D\mathrm{P}_{T_{D^\top \bar{x}}} D^\top (x^\star_{y,\lambda} - \bar{x})\right\| + \left\|\mathrm{P}_{S_{D^\top \bar{x}}} D^\top (x^\star_{y,\lambda} - \bar{x})\right\| \\
&= \left\|B^+_{\bar{x}V_{\bar{x}}} B_{\bar{x}} \left(\mathrm{Id}_n - D\mathrm{P}_{S_{D^\top \bar{x}}} D^\top\right) (x^\star_{y,\lambda} - \bar{x})\right\| + \left\|\mathrm{P}_{S_{D^\top \bar{x}}} D^\top (x^\star_{y,\lambda} - \bar{x})\right\| \\
&\leq \left\|B^+_{\bar{x}V_{\bar{x}}}\right\| \left\|B_{\bar{x}}(x^\star_{y,\lambda} - \bar{x})\right\| + \left(\left\|\mathrm{P}_{S_{D^\top \bar{x}}}\right\|_{\gamma_{\mathcal{C}} \to 2} + \left\|B^+_{\bar{x}V_{\bar{x}}}\right\| \left\|B_{\bar{x}} D_{S_{D^\top \bar{x}}}\right\|_{\gamma_{\mathcal{C}} \to 2}\right) \gamma_{\mathcal{C}} \left(\mathrm{P}_{S_{D^\top \bar{x}}} D^\top (x^\star_{y,\lambda} - \bar{x})\right) \\
&\leq \left\|B^+_{\bar{x}V_{\bar{x}}}\right\| \left\|B_{\bar{x}}(x^\star_{y,\lambda} - \bar{x})\right\| + \left(\left\|\mathrm{P}_{S_{D^\top \bar{x}}}\right\|_{\gamma_{\mathcal{C}} \to 2} + \left\|B^+_{\bar{x}V_{\bar{x}}}\right\| \left\|B_{\bar{x}} D_{S_{D^\top \bar{x}}}\right\|_{\gamma_{\mathcal{C}} \to 2}\right) \frac{D^v_R(x, \bar{x})}{1 - \sigma_{\mathcal{C}}\left(w_{S_{D^\top \bar{x}}}\right)},
\end{aligned}
$$

where we also used coercivity of $\gamma_{\mathcal{C}}$. Let

$$
\alpha = \frac{\left\|\mathrm{P}_{S_{D^\top \bar{x}}}\right\|_{\gamma_{\mathcal{C}} \to 2} + \left\|B^+_{\bar{x}V_{\bar{x}}}\right\| \left\|B_{\bar{x}} D_{S_{D^\top \bar{x}}}\right\|_{\gamma_{\mathcal{C}} \to 2}}{1 - \sigma_{\mathcal{C}}\left(w_{S_{D^\top \bar{x}}}\right)}.
$$

Thus

$$
\begin{aligned}
\left\|x^\star_{y,\lambda} - \bar{x}\right\| &\leq \left\|B^+_{\bar{x}V_{\bar{x}}}\right\| \left\||Ax^\star_{y,\lambda} - A\bar{x}|^2 + (|Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2)\right\| + \alpha D^v_R(x, \bar{x}) \\
&\leq \left\|B^+_{\bar{x}V_{\bar{x}}}\right\| \left(\|A\|^2 \left\|x^\star_{y,\lambda} - \bar{x}\right\|^2 + \left\||Ax^\star_{y,\lambda}|^2 - |A\bar{x}|^2\right\|\right) + \alpha D^v_R(x, \bar{x}).
\end{aligned}
$$

Inserting the bounds in (6.4.6) and rearranging, we get

$$
\left\|x^\star_{y,\lambda} - \bar{x}\right\| \leq b\sigma + a \left\|x^\star_{y,\lambda} - \bar{x}\right\|^2,
$$

where

$$
\begin{aligned}
a &= \frac{3}{2} \left\|B^+_{\bar{x}V_{\bar{x}}}\right\| \|A\|^2 + \alpha \|A\|^2 \frac{\|q\|}{2} \\
b &= \left\|B^+_{\bar{x}V_{\bar{x}}}\right\| \left(\frac{c\|q\|}{2} + \left(2 + \frac{c}{2}\|q\|\right)\right) + \alpha \frac{\left(2 + \frac{c}{2}\|q\|\right)^2}{2c}.
\end{aligned}
$$

When $-\bar{x}$ is the closest point, we argue similarly to get

$$
\left\|x^\star_{y,\lambda} + \bar{x}\right\| \leq b\sigma + a \left\|x^\star_{y,\lambda} + \bar{x}\right\|^2.
$$

Overall, we arrive at

$$
\mathrm{dist}(x^\star_{y,\lambda}, \overline{\mathcal{X}}) \leq b\sigma + a\mathrm{dist}(x^\star_{y,\lambda}, \overline{\mathcal{X}})^2.
$$

Solving the above inequality[6], we get that if

$$
\sigma \leq 1/(4ab),
$$

Then

$$
\mathrm{dist}(x^\star_{y,\lambda}, \overline{\mathcal{X}}) \leq \frac{1 - \sqrt{1 - 4ab\sigma}}{2a} \leq 2b\sigma.
$$

□

6.4.3 Convergence rate for Gaussian measurements

6.4.3.1 Construction of a "dual" certificate

The non-degenerate source condition (6.4.4) is a geometric condition, which is not easy to check in practice since exhibiting a valid non-degenerate dual certificate is not trivial for general A. We will

[6] Recall that $\mathrm{dist}(x^\star_{y,\lambda}, \overline{\mathcal{X}})$ vanishes as $\sigma \to 0$ thanks to Theorem 6.4.2.

now describe a particular construction of a good candidate (the so-called linearized pre-certificate). Moreover, when A is a Gaussian map, we will also provide sufficient bounds on m needed for conditions (6.4.4)-(6.4.5) to hold with overwhelming probability. In the sequel, the entries of A are i.i.d sampled from $\mathcal{N}(0, 1/m)$.

In the sequel, we will focus on the case where $D = \mathrm{Id}_n$. To lighten notation, we denote T and e the model parameters of R at $\bar{x}$.

Assume that (6.4.5). We define the vector

$$w \stackrel{\text{def}}{=} B_{\bar{x}}^\top \operatorname*{argmin}_{B_{\bar{x}}^\top q \in \mathrm{aff}(\partial R(\bar{x}))} \|q\| .$$

This is the minimal norm dual certificate. It can be easily shown, by definition of the model subspace T, that w can be equivalently expressed in closed form as

$$w = B_{\bar{x}}^\top B_{\bar{x}T}^{+,\top} e, \quad \text{where} \quad B_{\bar{x}T}^{+} = \left(B_{\bar{x}T}^\top B_{\bar{x}T}\right)^{-1} B_{\bar{x}T}^\top.$$

The goal is to investigate under which condition on m one can ensure that

$$\sigma_{\mathcal{C}}(w_S) < 1$$

with high probability.

Our approach is inspired by that of [57]. The key ingredient is the fact that, owing to the isotropy of the Gaussian ensemble, the actions on T and S are independent. However, unlike the linear case, in the phase retrieval problem, there is a major issue since $B_{\bar{x}}$ depends on $\bar{x}$. Therefore, our reasoning will also hold true for regularizers such that $\bar{x} \in T$. This covers the case of the ℓ_1 norm as well as the $\ell_1 - \ell_2$ norms. In this case, we can write

$$B_{\bar{x}} = \mathrm{diag}|A_T \bar{x}| A,$$

and thus

$$w = A^\top \mathrm{diag}|A_T \bar{x}|^2 A_T \left(A_T^\top \mathrm{diag}|A_T \bar{x}|^2 A_T\right)^{-1} e.$$

Define the vector

$$\eta \stackrel{\text{def}}{=} \mathrm{diag}|A_T \bar{x}|^2 A_T \left(A_T^\top \mathrm{diag}|A_T \bar{x}|^2 A_T\right)^{-1} e.$$

Clearly, isotropy of the Gaussian ensemble entails that η and A_S are independent, which allows us to infer the distribution of $A_S \eta$ with no knowledge of the values of A_T. Thus, for some $\tau > 0$ and $\nu \geq 1$

$$\Pr\left(\sigma_{\mathcal{C}}(w_S) \geq \nu\right) \leq \Pr\left(\sigma_{\mathcal{C}}(w_S) \geq \nu \middle| \|\eta\| \leq \tau\right) + \Pr\left(\|\eta\| \geq \tau\right). \tag{6.4.7}$$

The first term in this inequality will be bounded on a case-by-case basis (see the following sections) and uses the fact that conditionally on η, the entries of $w = A^\top \eta$ are *i.i.d* $\mathcal{N}(0, \|\eta\|^2 / m)$.

Let us consider the second term. We have the following.

Lemma 6.4.7. *If $m \geq C(\varrho) \log(m)$, on the same event we have*

$$\|\eta\| < \frac{1+\delta}{1-\varrho} \|e\|, \tag{6.4.8}$$

and

$$\|q\| < \frac{\sqrt{1+\delta}}{1-\varrho} \|e\| \frac{\sqrt{m}}{\|\bar{x}\|}, \tag{6.4.9}$$

with a probability at least $1 - \frac{6}{m^2} - e^{-\delta^2/2}$, where

$$q \stackrel{\text{def}}{=} \mathrm{diag}|A_T \bar{x}| A_T \left(A_T^\top \mathrm{diag}|A_T \bar{x}|^2 A_T\right)^{-1} e.$$

6.4.3.2 The ℓ_1 norm

In this case, $\sigma_{\mathcal{C}} = \|\cdot\|_\infty$, and where $s = |I|$, with $I = \mathrm{supp}(\bar{x})$. We have the following result.

Lemma 6.4.8. *Fix $\nu \in]0,1[$ and δ, ϱ (small enough), assume that the regularizer is the ℓ_1-norm we have,*

(i) if $m \geq C(\delta,\varrho)s\log(n)$ where $C(\delta,\varrho) = \frac{6(1+\delta)}{(1-\varrho)\nu^2}$ then

$$\left\|A_{I^c}^\top \eta\right\|_\infty < \nu,$$

with probability at least $1 - \frac{6}{m^2} - \frac{1}{n^2} - e^{-\delta^2/2}$.

(ii) if $m \geq C(\delta,\varrho)s\log^2(n)$ where $C(\delta,\varrho) = \frac{6(1+\delta)}{(1-\varrho)\nu^2}$ then we get the same result with probability at least $1 - \frac{6}{m^2} - \frac{1}{n^2} - m^{\delta-1}$.

We defer the proof to Section 6.5.1.1.

We now turn to state our convergence result for the ℓ_1-norm.

Proposition 6.4.9. *Fix $\nu \in]0,1[$ and δ, ϱ (small enough), and define $\zeta \overset{\text{def}}{=} \sqrt{\frac{1+\delta}{1-\varrho}}$. Let us consider the noisy phaseless measurements in (NoisyPR) with the Lasso penalty. Moreover let us choose $\lambda = c\sigma$, for some $c > 0$. If the entries of A are i.i.d sampled from $\mathcal{N}(0,1/m)$ with*

$$m \geq \frac{6\zeta^2}{\nu^2} s\log(n),$$

then with a probability at least $1 - \frac{1}{n^{\delta'-1}} - e^{-\delta^2/2} - \frac{6}{m^2}$ where $\delta' > 1$ is a constant, for σ small enough and any minimizer $x^\star_{y,\lambda}$, we have

$$\mathrm{dist}(x^\star_{y,\lambda}, \overline{\mathcal{X}}) \leq \left(2\zeta\left(c\zeta\sqrt{s}+2\right) + \left(\frac{c\zeta\sqrt{s}}{2}+2\right)^2 \frac{1+\zeta\sqrt{1+\delta}\left(\sqrt{m}+\sqrt{2\delta'\log(n)}\right)}{c(1-\nu)}\|\bar{x}\|\right)\sigma.$$

See Section 6.5.1.2 for the proof.

6.4.3.3 The $\ell_1 - \ell_2$ norm

We consider the group Lasso penalty. We recall that the number of blocks is L with equal size B and that $\bar{x}$ is s-block sparse. Let I be the (block) support of $\bar{x}$, *i.e.* $I = \mathrm{supp}_{\mathcal{B}}(\bar{x})$. From (6.2.10) we have that $\sigma_{\mathcal{C}}(v) = \max_{i\in I^c}\|v[b_i]\|_2$. Denote $\|v\|_{\infty,2} = \max_i \|v[b_i]\|_2$. We have the following.

Lemma 6.4.10. *Fix $\nu \in]0,1[$ and δ, ϱ (small enough), assume that the number of sample m is such that $m \geq \frac{1+\delta}{\left(\sqrt{\nu(1-\varrho)}-\sqrt{(1+\delta)s}\right)^2} s\left(\sqrt{B}+4\sqrt{\log(L)}\right)^2 + sB$ then*

$$\left\|A_{I^c}^\top \eta\right\|_{\infty,2} < \nu,$$

with probability at least $1 - L^{-7} - \frac{6}{m^2} - e^{-\frac{\delta^2}{2}}$.

The proof can be found in Section 6.5.2.1.

Proposition 6.4.11. *Fix $\nu \in]0,1[$ and δ, ϱ (small enough), and define $\zeta \overset{\text{def}}{=} \sqrt{\frac{1+\delta}{1-\varrho}}$. Let us consider the noisy phaseless measurements in (NoisyPR) with the group Lasso penalty. Moreover let us choose $\lambda = c\sigma$, for some $c > 0$. If the entries of A are i.i.d sampled from $\mathcal{N}(0,1/m)$ with*

$$m \geq \max\left(\frac{1+\delta}{\left(\sqrt{\nu(1-\varrho)}-\sqrt{(1+\delta)s}\right)^2}, \frac{1}{(1-\delta)^2}\right) s\left(\sqrt{B}+4\sqrt{\log(L)}\right)^2 + sB$$

then, with a probability at least $1 - L^{-7} - \frac{6}{m^2} - e^{-\frac{\delta^2}{2}} - L^{\delta'-1}$ *where* $\delta' > 1$ *is a constant, for* σ *small enough and any minimizer* $x^\star_{y,\lambda}$*, we have*

$$\operatorname{dist}(x^\star_{y,\lambda}, \overline{\mathcal{X}}) \leq \left(2\zeta \left(c\zeta\sqrt{s} + 2 \right) + \left(\frac{c\zeta\sqrt{s}}{2} + 2 \right)^2 \frac{1 + \left(\sqrt{m} + \sqrt{B} + \sqrt{2\delta \log(L)} \right) \delta \left\| \bar{x} \right\| \zeta}{c(1-\nu)} \right) \sigma.$$

The proof of this statement is similar to that of Proposition 6.4.9.

6.4.3.4 Symmetric strong gauge of a polytope

Here we suppose that

$$R = \gamma_{\mathcal{C}},$$

where $\mathcal{C}$ is a polytope. We use the shorthand notation $\mathcal{V}$ for the set of vertices of $\mathrm{P}_S\mathcal{C}$. We have the following.

Lemma 6.4.12. *Fix* $\nu \in]0,1[$ *and* δ, ϱ *(small enough),* $\bar{\alpha} \stackrel{\text{def}}{=} \frac{\|e\|^2 \max_{v \in \mathcal{V}_S} \|v\|^2}{\nu^2}$*. Let us assume that the number of samples* m *is such that* $m \geq 2\frac{(1+\delta)}{1-\varrho}\bar{\alpha}(1+\zeta)\log(|\mathcal{V}_S|)$ *where* $\zeta > 0$ *is a fixed numerical constant chosen arbitrary large, then*

$$\sigma_{\mathcal{C}}(A_S^\top \eta) < \nu,$$

with probability at least $1 - |\mathcal{V}_S|^{-\zeta} - \frac{6}{m^2} - e^{-\frac{\delta^2}{2}}$*.*

The proof is in Section 6.5.3.1.

As before, under this complexity bound, the conclusion of Theorem 6.4.4 holds with high probability. We do not restate it here for the sake of brevity.

6.5 Proofs for Section 6.4.3

6.5.0.1 Proof of Lemma 6.4.7

We have

$$\begin{aligned} \|\eta\|^2 &= \left\langle \operatorname{diag}|A_T\bar{x}|^2 A_T \left(A_T^\top \operatorname{diag}|A_T\bar{x}|^2 A_T \right)^{-1} e, \operatorname{diag}|A_T\bar{x}|^2 A_T \left(A_T^\top \operatorname{diag}|A_T\bar{x}|^2 A_T \right)^{-1} e \right\rangle, \\ &\leq \|A_T\bar{x}\|_\infty^4 \lambda_{\min}^2 \left(A_T^\top \operatorname{diag}|A_T\bar{x}|^2 A_T \right)^{-1} \|e\|^2. \end{aligned}$$

Thus,

$$\|\eta\| \leq \|A_T\bar{x}\|_\infty^2 \lambda_{\min} \left(A_T^\top \operatorname{diag}|A_T\bar{x}|^2 A_T \right)^{-1} \|e\|$$

By Lemma 6.6.1, we have

$$\|A_T\bar{x}\|_\infty^2 \leq \frac{1+\delta}{m} \|\bar{x}\|^2$$

with probability at least $1 - 2e^{-\frac{\delta^2}{2}}$. Observe also that

$$A_T^\top \operatorname{diag}|A_T\bar{x}|^2 A_T = \sum_{r=1}^m |\langle (a_r)_T, \bar{x} \rangle|^2 (a_r)_T (a_r)_T^\top.$$

From Lemma 6.6.2, as soon as $m \geq C(\varrho) d_T \log(m)$ we have

$$\lambda_{\min} \left(A_T^\top \operatorname{diag}|A_T\bar{x}|^2 A_T \right) \geq \frac{1-\varrho}{m} \|\bar{x}\|^2.$$

with a probability at least $1 - \frac{6}{m^2}$. Thus,

$$\Pr\left(\|\eta\| \geq \frac{1+\delta}{1-\varrho} \|e\| \right) \leq \frac{6}{m^2} + e^{-\frac{\delta^2}{2}}.$$

We have

$$\|q\|^2 = \left\langle \mathrm{diag}|A_T\bar{x}|A_T \left(A_T^\top \mathrm{diag}|A_T\bar{x}|^2 A_T\right)^{-1} e, \mathrm{diag}|A_T\bar{x}|A_T \left(A_T^\top \mathrm{diag}|A_T\bar{x}|A_T\right)^{-1} e \right\rangle,$$
$$\leq \|A_T\bar{x}\|_\infty^2 \lambda_{\min}^2 \left(A_T^\top \mathrm{diag}|A_T\bar{x}|^2 A_T\right)^{-1} \|e\|^2.$$

Thus,

$$\|q\| \leq \|A_T\bar{x}\|_\infty \lambda_{\min} \left(A_T^\top \mathrm{diag}|A_T\bar{x}|^2 A_T\right)^{-1}$$

By Lemma 6.6.1, we have

$$\|A_T\bar{x}\|_\infty \leq \|\bar{x}\| \sqrt{\frac{1+\delta}{m}}$$

with probability at least $1 - 2e^{-\frac{\delta^2}{2}}$. Observe also that

$$A_T^\top \mathrm{diag}|A_T\bar{x}|^2 A_T = \sum_{r=1}^m |\langle (a_r)_T, \bar{x}\rangle|^2 (a_r)_T (a_r)_T^\top.$$

From Lemma 6.6.2, as soon as $m \geq C(\varrho) d_T \log(m)$ we have

$$\lambda_{\min}\left(A_T^\top \mathrm{diag}|A_T\bar{x}|^2 A_T\right) \geq \frac{1-\varrho}{m} \|\bar{x}\|^2.$$

with a probability at least $1 - \frac{6}{m^2}$. Thus,

$$\Pr\left(\|q\| \geq \frac{\sqrt{1+\delta}}{1-\varrho} \|e\| \frac{\sqrt{m}}{\|\bar{x}\|}\right) \leq \frac{6}{m^2} + e^{-\frac{\delta^2}{2}}.$$

6.5.1 Proofs for the Lasso

6.5.1.1 Proof of Lemma 6.4.8

(i) The union bound and classical tail bounds of the Gaussian distribution give

$$\Pr\left(\left\|A_{I^c}^\top \eta\right\|_\infty \geq \nu \middle| \|\eta\| \leq \tau\right) \leq (n-s) \Pr\left(|Z| \geq \frac{\nu\sqrt{m}}{\tau}\right) \leq (n-s) e^{-\frac{m\nu^2}{2\tau^2}},$$

where $Z \sim \mathcal{N}(0,1)$. We invoke Lemma 6.4.7 and take $\tau = \sqrt{\frac{1+\delta}{1-\varrho}} s$ to get

$$\Pr\left(\left\|A_{I^c}^\top \eta\right\|_\infty \geq \nu \middle| \|\eta\| \leq \tau\right) \leq e^{-\frac{m(1-\varrho)\nu^2}{2s(1+\delta)} + \log(n-s)}.$$

As soon as $m \geq C(\delta,\varrho) s \log(n)$ we indeed have that $\left\|A_{I^c}^\top \eta\right\|_\infty < \nu$ with probability at least $1 - \frac{6}{m^2} - \frac{1}{n^2} - e^{-\delta/2}$ with $C(\delta,\varrho) = 6\frac{1+\delta}{(1-\varrho)\nu^2}$.

(ii) We can improve this probability by increasing the number of measurements. Indeed we also have by Lemma 6.6.1-(ii) that

$$\|Ax\|_\infty^2 \leq \frac{(1+\delta)\log(m)}{m} \|x\|^2.$$

Similar arguments the first case then yield

$$\Pr\left(\left\|A_{I^c}^\top \eta\right\|_\infty \geq \nu \middle| \|\eta\| \leq \tau\right) \leq e^{-\frac{-m(1-\varrho)\nu^2}{2s(1+\delta)\log(m)} + \log(n-s)}.$$

If $m \geq C(\delta,\varrho) s \log^2(n)$ we get that $\left\|A_{I^c}^\top \eta\right\|_\infty < \nu$ with probability at least $1 - \frac{6}{m^2} - \frac{1}{n^2} - m^{\delta-1}$ with $C(\delta,\varrho) = 6\frac{(1+\delta)}{(1-\varrho)\nu^2}$.

6.5.1.2 Proof of Proposition 6.4.9

The proof of this result involves applying the deterministic Theorem 6.4.4, along with additional arguments derived from the concentration properties of the random matrix A.

We have to bound the parameters b and α. We shall recall that,

$$b = \left\|B_T^+\right\| \left(\frac{c\,\|q\|}{2} + \left(2 + \frac{c}{2}\,\|q\|\right)\right) + \alpha \frac{\left(2 + \frac{c}{2}\,\|q\|\right)^2}{2c},$$

where

$$\alpha = \frac{\|\mathrm{P}_{S_{\bar{x}}}\|_{1\to 2} + \left\|B_{T_{\bar{x}}}^+\right\| \|B_{S_{\bar{x}}}\|_{1\to 2}}{1 - \left\|w_{S_{D^\top \bar{x}}}\right\|_\infty}.$$

6.5.2 Proofs for the group Lasso

6.5.2.1 Proof of Lemma 6.4.10

We denote $S_i = \mathrm{span}\{(a_j)_{\{j:\, j\in b_i\}}\}, i \in \mathrm{supp}_{\mathcal{B}}(x)^c$ then we have

$$\begin{aligned}
\Pr\left(\max_{i\in\mathrm{supp}_{\mathcal{B}}(\bar{x})^c} \left\|A_S^\top \eta[b_i]\right\| \geq \nu \middle| \|\eta\| \leq \tau\right) &= \Pr\left(\max_{i\in\mathrm{supp}_{\mathcal{B}}(\bar{x})^c} \|A_{S_i}\eta[b_i]\| \geq \nu \middle| \|\eta\| \leq \tau\right), \\
&\leq L \max_{i\in\mathrm{supp}_{\mathcal{B}}(\bar{x})^c} \Pr\left(\|A_{S_i}\eta[b_i]\| \geq \nu \middle| \|\eta\| \leq \tau\right), \\
&\leq L \max_{i\in\mathrm{supp}_{\mathcal{B}}(\bar{x})^c} \Pr\left(\|Z\| \geq \frac{\nu\sqrt{m}}{\tau}\right),
\end{aligned}$$

where Z is a standard Gaussian matrix of size $m \times B$, we have now to bound $\|Z\|$. By Proposition 2.6.8 and Gordon's Theorem [176, Theorem 5.32], we have

$$\begin{aligned}
\Pr\left(\|Z\| \geq \frac{\sqrt{m}}{\tau}\right) &= \Pr\left(\|Z\| - \mathbb{E}\left(\|Z\|\right) \geq \frac{\nu\sqrt{m}}{\tau} - \mathbb{E}\left(\|Z\|\right)\right), \\
&\leq \Pr\left(\|Z\| - \mathbb{E}\left(\|Z\|\right) \geq \frac{\nu\sqrt{m}}{\tau} - \sqrt{m} - \sqrt{B}\right), \\
&\leq \exp\left(-\frac{\left(\sqrt{m}\left(\frac{\nu}{\tau} - 1\right) - \sqrt{B}\right)^2}{2}\right),
\end{aligned}$$

Consequently,

$$\Pr\left(\max_{i\in\mathrm{supp}_{\mathcal{B}}(\bar{x})^c} \left\|A_S^\top \eta[b_i]\right\| \geq \nu \middle| \|\eta\| \leq \tau\right) \leq \exp\left(-\frac{\left(\sqrt{m}\left(\frac{\nu}{\tau} - 1\right) - \sqrt{B}\right)^2}{2} + \log(L)\right),$$

For $m \geq \frac{1+\delta}{\left(\sqrt{\nu(1-\varrho)} - \sqrt{(1+\delta)s}\right)^2} s\left(\sqrt{B} + 4\sqrt{\log(L)}\right)^2 + sB$, we get that $\max_{i\in\mathrm{supp}_{\mathcal{B}}(\bar{x})^c} \left\|A_S^\top \eta[b_i]\right\| < \nu$ with probability at least $1 - L^{-7} - \frac{6}{m^2} - e^{-\frac{\delta^2}{2}}$.

6.5.3 Proof for a symmetric strong gauge of a polytope

6.5.3.1 Proof of Lemma 6.4.12

We have that

$$\begin{aligned}\Pr\left(\sigma_{\mathcal{C}}(A_S^\top\eta)\geq\nu\middle|\,\|\eta\|\leq\tau\right)&=\Pr\left(\max_{v\in\mathcal{C}}\left\langle A_S^\top\eta,v\right\rangle\geq\nu\middle|\,\|\eta\|\leq\tau\right),\\&=\Pr\left(\max_{v\in\mathcal{V}_S}\langle A\eta,v\rangle\geq\nu\middle|\,\|\eta\|\leq\tau\right),\\&\leq|\mathcal{V}_S|\max_{v\in\mathcal{V}_S}\Pr\left(\langle Z\eta,v\rangle\geq\nu\sqrt{m}\middle|\,\|\eta\|\leq\tau\right),\end{aligned}$$

where Z is drawn from the standard Gaussian ensemble. Let us observe that $A\mapsto\langle A\eta,\nu\rangle$ is $\|\eta\|\,\|v\|$ −Lipschitz continuous function of matrices A considered as vector in $\mathbb{R}^{mn}$. From Proposition 2.6.8, we have that

$$\begin{aligned}\Pr\left(\sigma_{\mathcal{C}}(A_S^\top\eta)\geq\nu\middle|\,\|\eta\|\leq\tau\right)&\leq|\mathcal{V}_S|\max_{v\in\mathcal{V}_S}e^{-\frac{m\nu^2}{2\tau^2\|v\|^2}},\\&\leq|\mathcal{V}_S|\,e^{-\frac{m\nu^2}{2\tau^2\max_{v\in\mathcal{V}_S}\|v\|^2}},\\&=e^{-\frac{m\nu^2}{2\tau^2\max_{v\in\mathcal{V}_S}\|v\|^2}+\log(|\mathcal{V}_S|)}.\end{aligned}$$

we get that for the choice of $\bar{\alpha}$ and for $m\geq2\frac{(1+\delta)}{1-\varrho}\bar{\alpha}(1+\zeta)\log(|\mathcal{V}_S|)$ where $\zeta>0$ is a fixed numerical constant chosen arbitrary large, $\sigma_{\mathcal{C}}(A_S^\top\eta)<\nu$ with probability at least $1-|\mathcal{V}_S|^{-\zeta}-\frac{6}{m^2}-e^{-\frac{\delta^2}{2}}$.

6.6 Concentrations

Let us consider $T\subset\mathbb{R}^n$, denote $d=\dim(T)$ and consider A a $m\times d$ matrix whose entries are $i.i.d$ $\mathcal{N}(0,1/m)$. Throughout this section, we will see T through $\mathbb{R}^d$ since there exists an isometry between T and $\mathbb{R}^d$. We have the following concentrations.

Lemma 6.6.1. *Fix $\delta\in]0,1[$ we have,*

(i) for any $x\in T$

$$\|Ax\|_\infty^2\leq\frac{1+\delta}{m}\|x\|^2. \tag{6.6.1}$$

This happens with probability at least $1-e^{\frac{-\delta^2}{2}}$.

(ii) for any $x\in T$

$$\|Ax\|_\infty^2\leq\frac{(1+\delta)\log(m)}{m}\|x\|^2. \tag{6.6.2}$$

with probability $1-m^{\delta-1}$.

Proof. The proof comes easily from Lemma 3.6.4.

□

Lemma 6.6.2. *Fix $\varrho\in]0,1[$ (small enough) and choose $0<\bar{\varrho}<\frac{\varrho+3}{10\log(m)}$.*

(i) If the number of samples obeys $m\geq C(\varrho)d\log(d)$, for some sufficiently large $C(\varrho)>0$, we have

$$\left\|mA^\top\operatorname{diag}|A\bar{x}|^2A-\left(2\bar{x}\bar{x}^\top+\|\bar{x}\|^2\operatorname{Id}\right)\right\|\leq\varrho\|\bar{x}\|^2. \tag{6.6.3}$$

with a probability at least $1-5e^{-\zeta d}-\frac{4}{d^2}$ where ζ is a fixed numerical constant.

(ii) If the number of samples obeys $m\geq C(\bar{\varrho},\varrho)d\log(m)$, for some sufficiently large $C(\bar{\varrho},\varrho)>0$, (6.6.3) hold true with a probability at least $1-\frac{6}{m^2}$.

Proof. The proof of claim (i) is just an application of Lemma 3.6.2.

For the proof of claim (ii), we have to modify the choice of m in the different concentrations used in the proof of Lemma 3.6.2. We provide here a self-contained proof. We have to emphasize that showing (6.6.3) is similar to showing that

$$\left\| \frac{1}{m} \sum_{r=1}^{m} |\bar{a}_r[1]|^2 \bar{a}_r \bar{a}_r^\top - \left(2e_1 e_1^\top + \mathrm{Id} \right) \right\| \leq \bar{\varrho}, \tag{6.6.4}$$

where the entries of $\bar{a}_r$ are now standard Gaussian random variable and e_1 is a vector of the standard basis.

From symmetric arguments, showing (6.6.4) amounts to show that

$$V(v) \stackrel{\text{def}}{=} \left| \frac{1}{m} \sum_{r=1}^{m} |\bar{a}_r[1]|^2 |\bar{a}_r^\top v|^2 - \left(1 + 2v[1]^2 \right) \right| \leq \bar{\varrho}$$

for all $v \in \mathbb{S}^{d-1}$. The rest of the proof shows this claim.

Let $\widetilde{a}_r = (\bar{a}_r[2], \ldots, \bar{a}_r[d])$ and $\widetilde{v} = (v[2], \ldots, v[d])$. We rewrite

$$|\bar{a}_r^\top v|^2 = \left(\bar{a}_r[1] v[1] + \widetilde{a}_r^\top \widetilde{v} \right)^2 = \left(\bar{a}_r[1] v[1] \right)^2 + \left(\widetilde{a}_r^\top \widetilde{v} \right)^2 + 2 \bar{a}_r[1] v[1] \widetilde{a}_r^\top \widetilde{v}.$$

We plug this decomposition into $V(v)$ to get

$$\begin{aligned} V(v) \leq & \left| \frac{1}{m} \sum_{r=1}^{m} \bar{a}_r[1]^4 - 3 \right| v[1]^2 + \left| \frac{1}{m} \sum_{r=1}^{m} \bar{a}_r[1]^2 - 1 \right| \|\widetilde{v}\|^2 + 2 \left| \frac{1}{m} \sum_{r=1}^{m} |\bar{a}_r[1]|^3 v[1] \widetilde{a}_r^\top \widetilde{v} \right| \\ & + \left| \frac{1}{m} \sum_{r=1}^{m} \bar{a}_r[1]^2 \left(\widetilde{a}_r^\top \widetilde{v} - \|\widetilde{v}\|^2 \right) \right|. \end{aligned}$$

If $X \sim \mathcal{N}(0,1)$ we have $\mathbb{E}\left(X^{2p}\right) = \frac{(2p)!}{2^p p!}$ for $p \in \mathbb{N}$, and in particular $\mathbb{E}\left(X^2\right) = 1$ and $\mathbb{E}\left(X^4\right) = 3$. By the Tchebyshev's inequality and a union bound argument, $\forall \varepsilon > 0$, and a constant $C(\varepsilon) \approx \max\left(26, \frac{96}{\varepsilon^2}\right)$ such that when $m \geq C(\varepsilon)$ we have,

$$\frac{1}{m} \sum_{r=1}^{m} \left(\bar{a}_r[1]^4 - 3 \right) < \varepsilon, \quad \frac{1}{m} \sum_{r=1}^{m} \left(\bar{a}_r[1]^2 - 1 \right) < \varepsilon, \quad \frac{1}{m} \sum_{r=1}^{m} \bar{a}_r[1]^6 \leq 20$$
$$\text{and} \quad \max_{1 \leq r \leq m} |\bar{a}_r[1]| \leq \sqrt{10 \log m}.$$

Each of these events happens with probability at least $1 - \frac{1}{m^2}$, and thus their intersection occurs with probability at least $1 - \frac{4}{m^2}$. On this intersection event, we have

$$V(v) \leq \varepsilon + 2 \left| \frac{1}{m} \sum_{r=1}^{m} \bar{a}_r[1]^3 v[1] \widetilde{a}_r^\top \widetilde{v} \right| + \left| \frac{1}{m} \sum_{r=1}^{m} \bar{a}_r[1]^2 \left(\widetilde{a}_r^\top \widetilde{v} - \|\widetilde{v}\|^2 \right) \right|.$$

On the one hand, by a Hoeffding-type inequality (Proposition 2.6.6), we have

$$\forall \varrho' > 0, \quad \left| \frac{1}{m} \sum_{r=1}^{m} \bar{a}_r[1]^3 v[1] \widetilde{a}_r^\top \widetilde{v} \right| < \varrho' |v[1]| \, \|\widetilde{v}\|^2,$$

with a probability

$$p \geq 1 - e \exp\left(- \frac{c \varrho'^2 m^2}{d \sum_{r=1}^{m} a_r[1]^6} \right) \geq 1 - e \exp\left(- \frac{c \varrho'^2 m}{20 d} \right) \geq 1 - \exp\left(- \frac{2Cm}{d} \right),$$

where C is a constant that is large enough. When $m \geq \frac{1}{C} d \log(m)$ we get the bound with probability $p \geq 1 - \frac{1}{m^2}$. On the other hand, by Bernstein-type inequality (Proposition 2.6.7), we have

$$\forall \bar{\varrho} > 0, \quad \left| \frac{1}{m} \sum_{r=1}^{m} \bar{a}_r[1]^2 \left(\widetilde{a}_r^\top \widetilde{v} - \|\widetilde{v}\|^2 \right) \right| \leq \bar{\varrho} \, \|\widetilde{v}\|^2,$$

with probability

$$p' \geq 1 - \exp\left\{ - \min\left(\frac{\bar{\varrho}^2 m^2}{d \sum_{r=1}^{m} a_r[1]^4}; \frac{\bar{\varrho} m}{d \max\limits_{1 \leq r \leq m} a_r[1]^2} \right) \right\},$$
$$\geq 1 - \exp\left\{ - \min\left(\frac{\bar{\varrho}^2 m}{d(\varepsilon + 3)}; \frac{\bar{\varrho} m}{10 d \log(m)} \right) \right\},$$

For $\bar{\varrho} < \frac{\varepsilon+3}{10\log(m)}$, we get that $p' \geq 1 - \exp\left(-\frac{\bar{\varrho}^2 m}{d(\varepsilon+3)}\right) \geq 1 - \exp\left(-\frac{2C'm}{d}\right)$ for C' large enough. Thus, taking again $m \geq \frac{1}{C'} d \log(m)$ we get the bound with probability $p' \geq 1 - \frac{1}{m^2}$. Overall, for any $v \in \mathbb{S}^{n-1} \cap T$, we have with probability at least $1 - \frac{6}{m^2}$

$$V(v) \leq \varepsilon + \varrho' + 2\bar{\varrho}.$$

We conclude with a covering type argument which can be plugged into the sublinear term and we choose $m \geq C(\varrho, \bar{\varrho}) d \log(m)$ and observe that $\log(m) \geq \log(d)$. Therefore, choosing $\varrho = \varepsilon + \varrho' + 2\bar{\varrho}$, we get the claim. □

Chapter 7

Conclusion and Perspectives

7.1 Summary

This manuscript is concerned with the problem of phase retrieval. We have studied this problem from a theoretical and algorithmic point of view over the set of real vectors in finite dimension. We can split our analysis into two main parts: "Phase retrieval without regularization" and "Phase retrieval with regularization". In the first part, we have used a least-squared formulation to solve the problem. Due to the nonlinearity, the objective considered is nonconvex with a non-Lipschitz continuous gradient. Hence, we proposed to change the geometry to a Bregman-type one to solve the problem since this objective is smooth relative to an appropriate distance generating kernel. While in the second part, we have considered to minimize the sum of two functions. In fact, we add to the previous least-squared formulation a regularization term which promotes some prior knowledge about the object that we want to recover. In this setting, we have proposed and analyzed an algorithm that is suitable in this case: an inertial Bregman proximal gradient. Furthermore, we studied the noiseless and stability of the recovery for low complexity regularized phase retrieval.

We can sum up the main conclusions of our work in the following key points.

Phase retrieval without regularization.

(i) In the noiseless case, we can solve the phase retrieval problem based on the least square formulation using Bregman proximal gradient (aka mirror descent). For standard Gaussian measurement, if the number of measurements is sufficiently large, then for almost all initializers of the mirror descent with a given fixed step-size, we recover the true signal up to a global sign change with high probability. The convergence rate is linear and do not depend on the dimension of the problem. With a slightly smaller (polylog) number of measurements, we can afford to use a spectral initialization method to lie in a neighbourhood of the true vector and then recover the true one up to a global sign change. For the coded diffraction patterns model, we show local linear convergence to the true signal up to a global sign change with high probability for sufficiently large number of measurements, with the proviso that spectral initialization is used.

(ii) The mirror descent scheme is stable to small additive noise on the observations. Indeed, even in presence to noise bounded sequences converge to a critical point of the phase retrieval problem, and if the algorithm is well initialized and the noise is small enough, the critical point is near the true vector up to a global sign change. For standard Gaussian measurements, if the signal-to-noise ratio is large enough and the number of measurements is sufficiently large, almost all initializers globally converge to a global minimizer near the true vector (up to a global sign change). This sample complexity bound can be improved to (polylog) at the price of using a spectral method to provide a good initial guess. Furthermore, we have analyzed the geometry of the noisy objective function. When the number of measurements is sufficiently large and the

signal-to-noise ratio is large. In particular, the set of critical points of the objective function is reduced to the set of global minimizers and the set of strict saddle points.

Phase retrieval with regularization.

(iii) The inertial Bregman proximal method with the triangle scaling exponent property exhibits a similar behavior as Euclidean inertial forward-backward-type methods. Our results reveal that partial smoothness combined with a (generic) nondegeneracy condition allow our algorithm to identify activity in finite time and to enter a local (almost) linear regime restricted to a Riemannian manifold. Indeed, we have a global convergence of our iterates under the framework of the Kurdyka-Łojasiewicz property and a local linear convergence regime. In the case where the nonsmooth part vanishes, we show that for almost all initializers the generated sequences converge generically toward the set of critical points that are not strict saddle points, meaning that our algorithm escapes strict saddle points.

(iv) Noiseless and local stable recovery by low-complexity regularized phase retrieval is possible if the number of measurements is sufficiently large compared to the intrinsic complexity of the sought-after vector. This covers both sparse retrieval but also regularizers for which sour results are distinctly new.

7.2 Perspectives

Several extensions and directions are possible in the context of future work. Some of the most promising ones in our opinion are the following.

Structured measurement models. We have seen that the coded diffraction patterns model is challenging as it enjoys less randomness than the Gaussian model. We would like to understand this model more deeply from the theoretical point of view both for stability with and without regularization. Other questions remain open such as providing global recovery guarantees via mirror descent without spectral initialization. Such a result would be possible only if one can give a result on the landscape of the objective function in this setting. Stability results of mirror descent similar to those Chapter 4 for the Gaussian model are also lacking for the CDP model. More generally, many realistic phase-retrieval models still remain unexplored from a theoretical recovery viewpoint. This calls for more studies and necessitates a fruitful exchange between different disciplines, from applied physics to applied mathematics and computer science.

Regularized phase retrieval. Our guarantees in Chapter 6 are on global minimizers of the regularized phase retrieval minimization problem. To translate this in practice, one has to be ensured to have an algorithmic scheme that indeed converges near a global minimizer, or at least to have an initialization which provides such a good initial guess. This is for instance possible in the sparse case, but would degrade the sample complexity bound to the sparsity squared. It would then be interesting, and challenging, to design an algorithm or an initialization scheme for the general regularized case with sample complexity bounds that still scale linearly with the intrinsic dimension of the vector to recover. This is also in a close connection with studying the landscape of the regularized phase retrieval objective.

Escape property of the (I)BPG. It would be of strong interest to provide a generic escape property of Bregman-type proximal methods to minimize a smooth+nonsmooth objective under the "smooth adaptable" property which extends the Lipschitz continuity property of the gradient. Lipschitz continuity of the gradient of the smooth part, and more generally non-expansiveness, turns out to be instrumental to apply the centre stable manifold theorem in the Euclidean setting. The extension to the Bregman case is therefore an open challenging question.

Stochastic IBPG. We have only studied deterministic algorithms in this work where the whole gradient of the smooth part is computed at each iteration. However, since the latter has a finite sum structure, a stochastic version of (I)BPG can be designed where the gradient can be computed on small batch at each iteration. Studying the recovery guarantees for these stochastic schemes is an interesting and promising direction of furure work.

Machine learning for phase retrieval. Recently, neural networks based reconstruction algorithms have been applied to the phase retrieval problem with significant practical performance; see [71] for a recent overview. However, this comes with significant challenges including availability of training data, robustness issues, and the lack of theoretical reconstruction guarantees. This calls for more studies that we believe are worth investigating in the future.

List of Publications

Submitted or in preparation

(1) J-J. Godeme, J. Fadili, *Low Complexity Regularized Phase Retrieval.* In preparation.

(2) J.-J. Godeme, J. Fadili, *Inertial Bregman Proximal Gradient under Partial Smoothness.* In preparation.

(3) J.-J. Godeme, J. Fadili, M. Lequime, G. Soriano, C. Amra, and M. Zerrad, *Stable Phase Retrieval with Mirror Descent.* Submitted.

Journal Papers

(4) J.-J. Godeme, J. Fadili, X. Buet, M. Zerrad, M. Lequime, and C. Amra, *Provable Phase Retrieval with Mirror Descent,* SIAM J. Imaging Sci., 16(3):1106–1141, September 2023.

(5) X. Buet, M. Zerrad, M. Lequime, G. Soriano, J.-J. Godeme, J. Fadili, and C. Amra, *Instantaneous measurement of surface roughness spectra using white-light scattering projected on a spectrometer,.* Appl. Opt., AO, 62(7):B164–B169, March 2023.

(6) X. Buet, M. Zerrad, M. Lequime, G. Soriano, J.-J. Godeme, J. Fadili, and C. Amra, *Immediate and one-point roughness measurements using spectrally shaped light* Opt. Express, 30(10):16078–16093, May 2022.

Conference Papers

(7) J.-J. Godeme, M.J. Fadili, X. Buet, M. Zerrad, M. Lequime and C. Amra, *Reconstruction de Phase Garantie par Descente Miroir*, In 28th GRETSI Symposium on Signal and Image Processing, Nancy, 2022.

List of Notations

General definitions

$\mathbb{R}$: the set of real numbers
$\mathbb{R}_+$: nonnegative real numbers
$\mathbb{R}_{++}$: positive real numbers
$\overline{\mathbb{R}}$: $]-\infty, +\infty[\cup\{+\infty\}$, the extended real values
ℓ_+^1: nonnegative summable sequence
$\mathbb{N}$: set of nonnegative integers
$\mathbb{N}_+$: set of positive integers
$\mathbb{R}^n, \mathbb{R}^m$: finite dimensional real Euclidean spaces
Id: identity operator or $\mathbb{R}^n$
e: vector of all 1s

Set related

$\mathcal{S}$: a convex (often compact) set
$\iota_{\mathcal{S}}(\cdot)$: indicator function for the set $\mathcal{S}$
$\sigma_{\mathcal{S}}(\cdot)$: support function of the set $\mathcal{S}$
$\mathrm{P}_{\mathcal{S}}(\cdot)$: projection operator onto $\mathcal{S}$
$\mathrm{int}\mathcal{S}$: interior of $\mathcal{S}$
$\bar{\mathcal{S}}$: closure of $\mathcal{S}$
$\mathrm{ri}(\mathcal{S})$: relative interior of $\mathcal{S}$
$\mathrm{aff}(\mathcal{S})$: smallest affine subspace that contains $\mathcal{S}$, a.k.a. affine hull of $\mathcal{S}$
$\mathrm{par}(\mathcal{S})$: the subspace parallel to $\mathcal{S}$
A^{-1}: inverse of A
$\mathrm{dom}(A)$: domain of A
$\mathrm{ran}(A)$: range of A
argmin: the set of minimizing arguments
$B(x, r)$: a ball centered at x with radius $r > 0$

Function related

$\Gamma_0(\mathbb{R}^n)$: the set of proper convex and lower semi-continuous functions on a $\mathbb{R}^n$.
$\mathrm{dom}(R)$: domain of R
R^*: Fenchel conjugate of R
∇F: gradient of F
$\mathrm{prox}_{\gamma R}$: proximal operator of R with $\gamma > 0$
∂R: subdifferential of function R
$(\gamma_k)_{k\in\mathbb{N}}$: a sequence indexed by k
D_R^v: Bregman divergence of R associated to $v \in \partial R$
$\mathbb{E}[x]$: total expectation of the random variable x
$\mathbb{P}$: a probability measure

List of Figures

Bibliography

[1] E. J. Akutowicz. On the determination of the phase of a Fourier integral, I. *Transactions of the American Mathematical Society*, 83(1):179, September 1956.

[2] E. J. Akutowicz. On the determination of the phase of a Fourier integral, II. *Proceedings of the American Mathematical Society*, 8(2):234, April 1957.

[3] F. Alvarez. On the Minimizing Property of a Second Order Dissipative System in Hilbert Spaces. *SIAM J. Control Optim.*, 38(4):1102–1119, January 2000.

[4] F. Alvarez and H. Attouch. An Inertial Proximal Method for Maximal Monotone Operators via Discretization of a Nonlinear Oscillator with Damping. *Set-Valued Analysis*, 9(1):3–11, March 2001.

[5] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference*, 3(3):224–294, September 2014.

[6] C. Amra, M. Lequime, and M. Zerrad. *Electromagnetic Optics of Thin-Film Coatings: Light Scattering, Giant Field Enhancement, and Planar Microcavities*. Cambridge University Press, Cambridge, 2021.

[7] C. Amra, M. Zerrad, S. Liukaityte, and M. Lequime. Instantaneous one-angle white-light scatterometer. *Opt. Express, OE*, 26(1):204–219, January 2018.

[8] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Math. Program.*, 137(1-2):91–129, February 2013.

[9] H. Attouch and J. Peypouquet. The Rate of Convergence of Nesterov's Accelerated Forward-Backward Method is Actually Faster Than $1/k^2$. *SIAM J. Optim.*, 26(3):1824–1834, January 2016.

[10] H. Attouch, J. Peypouquet, and P. Redont. A Dynamical Approach to an Inertial Forward-Backward Algorithm for Convex Minimization. *SIAM J. Optim.*, 24(1):232–256, January 2014.

[11] J.-P. Aubin and I. Ekeland. *Applied Nonlinear Analysis*. Elsevier, 1984.

[12] A. Auslender and M. Teboulle. Asymptotic Cones and Functions in Optimization and Variational Inequalities. *Springer Monographs in Mathematics*, pages 25–80, 2003.

[13] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.*, 16(3):697–725, January 2006.

[14] S. Bahmani and J. Romberg. Efficient Compressive Phase Retrieval with Constrained Sensing Vectors. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[15] R. Balan. Reconstruction of signals from magnitudes of redundant representations: The complex case. *Found Comput Math*, 16(3):677–721, June 2016.

[16] R. Balan, P. Casazza, and D. Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, May 2006.

[17] A. S. Bandeira, J. Cahill, D. G. Mixon, and A. A. Nelson. Saving phase: Injectivity and stability for phase retrieval. *arXiv:1302.4618 [math]*, October 2013.

[18] A. S. Bandeira and D. G. Mixon. Near-optimal phase retrieval of sparse vectors. In *SPIE Proceedings*. SPIE, September 2013.

[19] R. Barakat and G. Newsam. Algorithms for reconstruction of partially known, band-limited Fourier-transform pairs from noisy data. *J. Opt. Soc. Am. A, JOSAA*, 2(11):2027–2039, November 1985.

[20] H. H. Bauschke, J. Bolte, J. Chen, M. Teboulle, and X. Wang. On linear convergence of non-euclidean gradient methods without strong convexity and Lipschitz gradient continuity. *Journal of Optimization Theory and Applications*, 182(3):1068–1087, 2019.

[21] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, page 20, 2016.

[22] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

[23] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization. *J. Opt. Soc. Am. A*, 19(7):1334, July 2002.

[24] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Finding best approximation pairs relative to two closed convex sets in Hilbert spaces. *J. Approx. Theory*, 127:178–192, 2004.

[25] H. H. Bauschke, D. R. Luke, H. M. Phan, and X. Wang. Restricted normal cones and the method of alternating projections: applications. *Set-Valued and Variational Analysis*, 21:475–501, 2013.

[26] Heinz H. Bauschke, Jonathan M. Borwein, and Patrick L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, 42(2):596–636, 2003.

[27] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sci.*, 2(1):183–202, January 2009.

[28] R. Beinert and G. Plonka. Ambiguities in one-dimensional discrete phase retrieval from Fourier magnitudes. *J. Fourier Ana. App.*, 21(6):1169–1198, 2015.

[29] R. Beinert and M. Quellmalz. Total Variation-Based Reconstruction and Phase Retrieval for Diffraction Tomography. *SIAM Journal on Imaging Sciences*, 15:1373–1399, September 2022.

[30] R. Beinert and M. Quellmalz. Total Variation-Based Reconstruction and Phase Retrieval for Diffraction Tomography with an Arbitrarily Moving Object. *PAMM*, 22(1):e202200135, 2023.

[31] A. Ben-Tal and A. S. Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. MPS-SIAM series on optimization. Society for Industrial and Applied Mathematics, 2001.

[32] T. Bendory, R. Beinert, and Y. C. Eldar. Fourier phase retrieval: Uniqueness and algorithms. In Holger Boche, Giuseppe Caire, Robert Calderbank, Maximilian März, Gitta Kutyniok, and Rudolf Mathar, editors, *Compressed Sensing and its Applications*, Applied and Numerical Harmonic Analysis, pages 55–91. Birkhäuser, 2017.

[33] P. Bianchi, W. Hachem, and S. Schechtman. Stochastic Subgradient Descent Escapes Active Strict Saddles on Weakly Convex Functions. *Mathematics of OR*, September 2023.

[34] B. Birnbaum, N. R. Devanur, and L. Xiao. Distributed algorithms via gradient descent for fisher markets. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 127–136, 2011.

[35] J. Bolte, A. Daniilidis, and A. Lewis. The łojasiewicz Inequality for Nonsmooth Subanalytic Functions with Applications to Subgradient Dynamical Systems. *SIAM J. Optim.*, 17(4):1205–1223, January 2007.

[36] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[37] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of łojasiewicz inequalities: Subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc.*, 362(06):3319–3363, December 2009.

[38] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2):459–494, August 2014.

[39] J. Bolte, S. Sabach, and M. Teboulle. Nonconvex lagrangian-based optimization: Monitoring schemes and global convergence. *Mathematics of Operations Research*, 43(4):1210–1232, 2018.

[40] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM J. Optim.*, 28(3):2131–2151, 2018.

[41] R. I. Bot, E. Robert Csetnek, and S. C. László. An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions. *EURO Journal on Computational Optimization*, 4(1):3–25, 2016.

[42] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, February 2013.

[43] N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[44] O. Brandière and M. Duflo. Les algorithmes stochastiques contournent-ils les pièges ? *Annales de l'I.H.P. Probabilités et statistiques*, 32:395–427, 1996.

[45] X. Buet, M. Zerrad, M. Lequime, G. Soriano, J.-J. Godeme, J. Fadili, and C. Amra. Immediate and one-point roughness measurements using spectrally shaped light. *Opt. Express*, 30(10):16078–16093, May 2022.

[46] X. Buet, M. Zerrad, M. Lequime, G. Soriano, J.-J. Godeme, J. Fadili, and C. Amra. Instantaneous measurement of surface roughness spectra using white-light scattering projected on a spectrometer. *Appl. Opt., AO*, 62(7):B164–B169, March 2023.

[47] P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, Berlin, Heidelberg, 2011.

[48] J. V. Burke and J. J. More. On the Identification of Active Constraints. *SIAM Journal on Numerical Analysis*, 25(5):1197–1211, 1988.

[49] J.-F. Cai and W. Xu. Guarantees of total variation minimization for signal recovery. *Information and Inference: A Journal of the IMA*, 4(4):328–353, December 2015.

[50] T. T. Cai, X. Li, and Z. Ma. Optimal Rates of Convergence for Noisy Sparse Phase Retrieval Via Thresholded Wirtinger Flow. *The Annals of Statistics*, 44(5):2221–2251, 2016.

[51] E. Candès and X. Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Found. Comput. Math.*, 2014.

[52] E. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, September 2015.

[53] E. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Trans. Inform. Theory*, 61(4):1985–2007, 2015.

[54] E. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98:925–936, 2010.

[55] E. Candés, T. Strohmer, and V. Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[56] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[57] E. J. Candès and B. Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1–2):577–589, 2013.

[58] A. Chambolle and Ch. Dossal. On the Convergence of the Iterates of the "Fast Iterative Shrinkage/Thresholding Algorithm". *J Optim Theory Appl*, 166(3):968–982, September 2015.

[59] Antonin Chambolle and Jérôme Darbon. A parametric maximum flow approach for discrete total variation regularization. In *Image Processing and Analysis with Graphs*. CRC Press, 2012.

[60] R. Chandra, Z. Zhong, J. Hontz, V. McCulloch, C. Studer, and T. Goldstein. Phasepack: A phase retrieval library. *Asilomar Conference on Signals, Systems, and Computers*, 2017.

[61] V. Chandrasekaran, B. Recht, A. Parrilo, P., and S. Willsky, A. The Convex Geometry of Linear Inverse Problems. *Found Comput Math*, 12(6):805–849, December 2012.

[62] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.*, 3(3):538–543, August 1993.

[63] Y. Chen and E. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Comm. Pure Appl. Math.*, 70(5):822–883, May 2017.

[64] Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Math. Program.*, 176(1-2):5–37, July 2019.

[65] T. R. Crimmins and J. R. Fienup. Ambiguity of phase retrieval for functions with disconnected support. *J. Opt. Soc. Am.*, 71(8):1026, August 1981.

[66] T. R. Crimmins and J. R. Fienup. Uniqueness of phase retrieval for functions with sufficiently disconnected support. *J. Opt. Soc. Am.*, 73(2):218, February 1983.

[67] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM J. Numer. Anal.*, 7:1–4, 1970.

[68] D. Davis and D. Drusvyatskiy. Proximal Methods Avoid Active Strict Saddles of Weakly Convex Functions. *Found Comput Math*, 22(2):561–606, April 2022.

[69] D. Davis, D. Drusvyatskiy, and C. Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis*, 40(4):2652–2695, October 2020.

[70] L. Demanet and P. Hand. Stable Optimizationless Recovery from Phaseless Linear Measurements. *MIT web domain*, November 2013.

[71] Jonathan Dong, Lorenzo Valzania, Antoine Maillard, Thanh-an Pham, Sylvain Gigan, and Michael Unser. Phase retrieval: From computational imaging to machine learning: A tutorial. *IEEE Signal Processing Magazine*, 40(1):45–57, 2023.

[72] R.-A. Dragomir, A. B. Taylor, A. d'Aspremont, and J. Bolte. Optimal complexity and certification of Bregman first-order methods. *Math. Program.*, 194(1):41–83, July 2022.

[73] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Transversality and alternating projections for nonconvex sets. *Found. Comput. Math.*, 15(6):1637–1651, 2015.

[74] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Generic Minimizing Behavior in Semialgebraic Optimization. *SIAM J. Optim.*, 26(1):513–534, January 2016.

[75] D. Drusvyatskiy and A. S. Lewis. Optimality, identifiability, and sensitivity. *Math. Program.*, 147(1):467–498, October 2014.

[76] J. C. Dunn. On the convergence of projected gradient processes to singular critical points. *J Optim Theory Appl*, 55(2):203–216, November 1987.

[77] Y. C. Eldar and Mendelson S. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[78] A. Fannjiang and T. Strohmer. The numerics of phase retrieval. *Acta Numerica*, 29:125–228, May 2020.

[79] J. R. Fienup. Phase retrieval algorithms: a comparison. *Appl. Opt.*, 21(15):2758, August 1982.

[80] B. Gao, Y. Wang, and Z. Xu. Stable Signal Recovery from Phaseless Measurements. *J Fourier Anal Appl*, 22(4):787–808, August 2016.

[81] M. Genzel, M. März, and R. Seidel. Compressed Sensing with 1D Total Variation: Breaking Sample Complexity Barriers via Non-Uniform Recovery. *Information and Inference: A Journal of the IMA*, 11:203–250, March 2022.

[82] R. Gerchberg and W. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35(2):237, 1972.

[83] J.-J. Godeme, J. Fadili, X. Buet, M. Zerrad, M. Lequime, and C. Amra. Provable phase retrieval with mirror descent. *SIAM J. Imaging Sci.*, 16(3):1106–1141, September 2023.

[84] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[85] Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in ?n. In J. Lindenstrauss and V. D. Milman, editors, *Geometric Aspects of Functional Analysis*, Lecture Notes in Mathematics, pages 84–106, Berlin, Heidelberg, 1988. Springer.

[86] X. Goudou and J. Munier. The gradient and heavy ball with friction dynamical systems: the quasiconvex case. *Math. Program.*, 116(1):173–191, January 2009.

[87] D. Gross, F. Krahmer, and R. Kueng. Improved recovery guarantees for phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 42(1):37–64, January 2017.

[88] P. Hand and V. Voroninski. Compressed Sensing from Phaseless Gaussian Measurements via Linear Programming in the Natural Parameter Space. *ArXiv*, November 2016.

[89] F. Hanzely, P. Richtárik, and L. Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Comput Optim Appl*, 79(2):405–440, June 2021.

[90] W. L. Hare. Identifying Active Manifolds in Regularization Problems. In H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Springer Optimization and Its Applications, pages 261–271. Springer, New York, NY, 2011.

[91] W. L. Hare and A. S. Lewis. Identifying Active Constraints via Partial Smoothness and Prox-Regularity. *Journal of Convex Analysis*, 2004.

[92] W. L. Hare and A. S. Lewis. Identifying Active Manifolds. *Algorithms Operations Research*, pages 75–82, 2007.

[93] M. Hayes. The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30(2):140–154, April 1982.

[94] R. Hesse and D. R. Luke. Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. *SIAM J. Optim.*, 23(4):2397–2419, 2013.

[95] L. T. K. Hien, N. Gillis, and P. Patrinos. Inertial block proximal methods for non-convex non-smooth optimization, 2020.

[96] A. D. Ioffe. *Variational Analysis of Regular Mappings*. Springer Monographs in Mathematics. Springer International Publishing, Cham, 2017.

[97] K. Jaganathan, Eldar Y. C., and B. Hassibi. Phase retrieval: An overview of recent developments. In A. Stern, editor, *Optical Compressive Imaging*. CRC Press, 2016.

[98] K. Jaganathan, S. Oymak, and B. Hassibi. Recovery of sparse 1-D signals from the magnitudes of their Fourier transform. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 1473–1477, July 2012. ISSN: 2157-8117.

[99] K. Jaganathan, S. Oymak, and B. Hassibi. Sparse phase retrieval: Convex algorithms and limitations. In *2013 IEEE International Symposium on Information Theory*, pages 1022–1026, July 2013. ISSN: 2157-8117.

[100] K. Jaganathan, S. Oymak, and B. Hassibi. Sparse Phase Retrieval: Uniqueness Guarantees and Recovery Algorithms. *IEEE Transactions on Signal Processing*, 65(9):2402–2410, May 2017.

[101] G. Jagatap and C. Hegde. Fast, Sample-Efficient Algorithms for Structured Phase Retrieval. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[102] F. Krahmer and D. Stöger. Complex Phase Retrieval from Subgaussian Measurements. *J. Fourier Anal Appl*, 26(6):89, November 2020.

[103] F. Krahmer and R. Ward. Stable and Robust Sampling Strategies for Compressive Imaging. *IEEE Transactions on Image Processing*, 23(2):612–622, February 2014.

[104] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998.

[105] M. Laghdir and M. Volle. A general formula for the horizon function of a convex composite function. *Archiv der Mathematik*, 73(4):291–302, Oct 1999.

[106] E. Laude, P. Ochs, and D. Cremers. Bregman proximal mappings and Bregman–Moreau envelopes under relative prox-regularity. *J Optim Theory Appl*, 184(3):724–761, March 2020.

[107] G. Lecué and S. Mendelson. Minimax rate of convergence and the performance of empirical risk minimization in phase recovery. *Electronic Journal of Probability*, 20:1–29, January 2015.

[108] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, Rhode Island, February 2005.

[109] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, I. M. Jordan, and B. Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176(1):311–337, 2019.

[110] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht. First-order methods almost always avoid strict saddle points. *Math. Program.*, 176(1-2):311–337, July 2019.

[111] A. S. Lewis. Active Sets, Nonsmoothness, and Sensitivity. *SIAM J. Optim.*, 13(3):702–725, January 2002.

[112] A. S. Lewis, D. R. Luke, , and J. Malick. Local linear convergence of alternating and averaged projections. *Found. Comput. Math.*, 9(4):485–513, 2009.

[113] A. S. Lewis and J. Malick. Alternating projections on manifolds. *Math. Oper. Res.*, 33:216–234, 2008.

[114] X. Li and V. Voroninski. Sparse Signal Recovery from Quadratic Measurements via Convex Programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, January 2013.

[115] J. Liang. *Convergence rates of first-order operator splitting methods*. PhD thesis, Université de Caen, 2016.

[116] J. Liang, J. Fadili, and G. Peyré. Local linear convergence of forward–backward under partial smoothness. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 1970–1978, Cambridge, MA, USA, December 2014. MIT Press.

[117] J. Liang, J. Fadili, and G. Peyré. A Multi-step Inertial Forward-Backward Splitting Method for Non-convex Optimization. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[118] J. Liang, J. Fadili, and G. Peyré. Activity Identification and Local Linear Convergence of Forward–Backward-type Methods. *SIAM J. Optim.*, 27(1):408–437, January 2017.

[119] J. Liang, J. Fadili, and G. Peyré. Local linear convergence analysis of Primal–Dual splitting methods. *Optimization*, 67(6):821–853, June 2018.

[120] J. Liang, J. Fadili, G. Peyré, and R. Luke. Activity Identification and Local Linear Convergence of Douglas–Rachford/ADMM under Partial Smoothness. In J.-F. Aujol, M. Nikolova, and N. Papadakis, editors, *Scale Space and Variational Methods in Computer Vision*, Lecture Notes in Computer Science, pages 642–653. Springer International Publishing, 2015.

[121] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles*, pages 87–89. Editions du Centre National de la Recherche Scientifique, 1963.

[122] S. Łojasiewicz. Ensembles semi-analytiques. *Lectures Notes IHES (Bures-sur-Yvette)*, 1965.

[123] D. A. Lorenz and T. Pock. An Inertial Forward-Backward Algorithm for Monotone Inclusions. *J Math Imaging Vis*, 51(2):311–325, February 2015.

[124] H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[125] D. R. Luke. Finding best approximation pairs relative to a convex and a prox-regular set in Hilbert space. *SIAM J. Optim.*, 19(2):714–739, 2008.

[126] D. R. Luke. Local linear convergence of approximate projections onto regularized sets. *Nonlinear Anal.*, 75:1531–1546, 2012.

[127] D. R. Luke. Phase Retrieval, What's New? *SIAG/OPT Views and News*, 25(1):1–6, 2017.

[128] D. R. Luke and A.-L. Martins. Convergence analysis of the relaxed douglas–rachford algorithm. *SIAM Journal on Optimization*, 30(1):542–584, 2020.

[129] D. R. Luke, N. H. Thao, and M. K. Tam. Quantitative convergence analysis of iterated expansive, set-valued mappings. *Mathematics of Operations Research*, 43(4):1143–1176, 2018.

[130] T. D. Luu, J. Fadili, and C. Chesneau. Sharp oracle inequalities for low-complexity priors. *Ann Inst Stat Math*, 72(2):353–397, April 2020.

[131] AkÃSakaya M. and Tarokh V. New conditions for sparse phase retrieval. *ArXiv*, abs/1310.1351, 2013.

[132] Huang M. and Xu Z. Performance bound of the intensity-based model for noisy phase retrieval, 2021.

[133] A. D. McRae, J. Romberg, and M. A. Davenport. Optimal Convex Lifted Sparse Phase Retrieval and PCA With an Atomic Matrix Norm Regularizer. *IEEE Transactions on Information Theory*, 69(3):1866–1882, March 2023.

[134] J. Miao, P. Charalambous, J. Kirz, and D. Sayre. Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400:342–344, 1999.

[135] A. Moudafi and M. Oliny. Convergence of a splitting inertial proximal method for monotone operators. *Journal of Computational and Applied Mathematics*, 155(2), June 2003.

[136] M. C. Mukkamala, P. Ochs, T. Pock, and S. Sabach. Convex-concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization. *SIAM Journal on Mathematics of Data Science*, 2(3):658–682, January 2020.

[137] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, May 2009.

[138] D. Needell and R. Ward. Near-Optimal Compressed Sensing Guarantees for Total Variation Minimization. *IEEE Transactions on Image Processing*, 22(10):3941–3949, October 2013.

[139] D. Needell and R. Ward. Stable Image Reconstruction Using Total Variation Minimization. *SIAM J. Imaging Sci.*, 6(2):1035–1058, January 2013.

[140] Y. Nesterov. A method for solving the convex programming problem with convergence rate O(1/k^2). *Proceedings of the USSR Academy of Sciences*, 1983.

[141] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. *IEEE Transactions on Signal Processing*, 63(18):4814–4826, 2015.

[142] D. Noll and A. Rondepierre. On local convergence of the method of alternating projections. *Found. Comput. Math.*, 16(2):425–455, 2016.

[143] H. Ohlsson and Y. C. Eldar. On conditions for uniqueness in sparse phase retrieval. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1841–1845, May 2014.

[144] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously Structured Models With Application to Sparse and Low-Rank Matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908, May 2015.

[145] S. Oymak, C. Thrampoulidis, and B. Hassibi. The squared-error of generalized LASSO: A precise analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1002–1009, October 2013.

[146] I. Panageas and G. Piliouras. Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions, 2017.

[147] E. Pauwels, A. Beck, Y. C. Eldar, and S. Sabach. On fienup methods for sparse phase retrieval. *IEEE Transactions on Signal Processing*, 66(4):982–991, February 2018.

[148] R. Pedarsani, D. Yin, K. Lee, and K. Ramchandran. PhaseCode: Fast and Efficient Compressive Phase Retrieval Based on Sparse-Graph Codes. *IEEE Transactions on Information Theory*, 63(6):3663–3691, June 2017.

[149] R. Pemantle. Nonconvergence to Unstable Points in Urn Models and Stochastic Approximations. *The Annals of Probability*, 18(2):698–712, April 1990.

[150] H. Phan. Linear convergence of the Douglas-Rachford method for two closed sets. *Optimization*, 65:369–385, 2016.

[151] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, January 1964.

[152] C. Poon. On the Role of Total Variation in Compressed Sensing. *SIAM J. Imaging Sci.*, 8(1):682–720, January 2015.

[153] N. Rao, B. Recht, and R. Nowak. Universal Measurement Bounds for Structured Sparse Signal Recovery. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 942–950, March 2012.

[154] N. S. Rao, B. Recht, and R. Nowak. Signal Recovery in Unions of Subspaces with Applications to Compressive Imaging. *arXiv: Machine Learning*, September 2012.

[155] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[156] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[157] H. Sahinoglou and S. Cabrera. On phase retrieval of finite-length sequences using the initial time sample. *IEEE Transactions on Circuits and Systems*, 38(5):954–958, 1991.

[158] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational Methods in Imaging*, volume 167 of *Applied Mathematical Sciences*. Springer, New York, NY, 2009. ISSN: 0066-5452.

[159] P. Schniter and S. Rangan. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055, 2015.

[160] A. Shapiro. Second order sensitivity analysis and asymptotic theory of parametrized nonlinear programs. *Mathematical Programming*, 33(3):280–299, December 1985.

[161] Y Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, May 2015.

[162] M. Shub. *Global Stability of Dynamical Systems*. Springer, 1987.

[163] A. Silveti-Falls, C. Molinari, and J. Fadili. A stochastic Bregman primal-dual splitting algorithm for composite optimization. *Pure and Applied Functional Analysis (special issue in honor of L. Bregman)*, 2022.

[164] M. Soltanolkotabi. *Algorithms and Theory for Clustering and Nonconvex Quadratic Programming*. PhD thesis, Stanford University, 2014.

[165] M. Soltanolkotabi. Structured Signal Recovery From Quadratic Measurements: Breaking Sample Complexity Barriers via Nonconvex Optimization. *IEEE Transactions on Information Theory*, 65(4):2374–2400, April 2019.

[166] J.-L. Starck, F. Murtagh, and J. Fadili. *Sparse Image and Signal Processing: Wavelets and Related Geometric Multiscale Analysis, Second Edition*. Cambridge University Press, 2 edition, 2015.

[167] G. Steidl, J. Weickert, T. Brox, P. Mrázek, and M. Welk. On the Equivalence of Soft Wavelet Shrinkage, Total Variation Diffusion, Total Variation Regularization, and SIDEs. *SIAM J. Numer. Anal.*, 42(2):686–713, January 2004.

[168] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *Found Comput Math*, 18(5):1131–1198, October 2018.

[169] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96, 2018.

[170] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems.* V. H. Winston and Sons, October 1977.

[171] J. A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. *Found Comput Math*, 12(4):389–434, August 2012.

[172] J. A. Tropp. Convex Recovery of a Structured Signal from Independent Random Linear Measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Birkhäuser, 2015.

[173] S. Vaiter, M. Golbabaee, J. Fadili, and G. Peyré. Model selection with low complexity priors. *Information and Inference: A Journal of the IMA*, 4(3):230–287, September 2015.

[174] S. Vaiter, G. Peyré, and J. Fadili. Low Complexity Regularization of Linear Inverse Problems. In Götz E. Pfander, editor, *Sampling Theory, a Renaissance: Compressive Sensing and Other Developments*, Applied and Numerical Harmonic Analysis, pages 103–153. Springer International Publishing, 2015.

[175] N. Vaswani. Non-convex structured phase retrieval. *arXiv:2006.13298 [cs, eess, math, stat]*, June 2020.

[176] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027 [cs, math]*, November 2011.

[177] J. Von Neumann. *Some matrix inequalities and metrization of matric space.* In: Collected Works, Vol. IV, Pergamon, Oxford, 1962.

[178] V. Voroninski and Z. Xu. A strong restricted isometry property, with an application to phaseless compressed sensing. *Applied and Computational Harmonic Analysis*, 40(2):386–395, March 2016.

[179] I. Waldspurger. Phase retrieval with random gaussian sensing vectors by alternating projections. *IEEE Transactions on Information Theory*, 64(5):3301–3312, May 2018.

[180] I. Waldspurger, A. d'Aspremont, and S. Mallat. Phase recovery, MaxCut and complex semidefinite programming. *Math. Program.*, 149(1):47–81, February 2015.

[181] A. Walther. The question of phase retrieval in optics. *Optica Acta: International Journal of Optics*, 10(1):41–49, January 1963.

[182] G. Wang, G. B. Giannakis, and Y. C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *arXiv:1605.08285 [cs, math, stat]*, August 2017.

[183] G. Wang, Georgios B. Giannakis, J. Chen, and M. Akçakaya. SPARTA: Sparse phase retrieval via Truncated Amplitude flow. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3974–3978, March 2017.

[184] Y. Wang and Z. Xu. Phase retrieval for sparse signals. *Applied and Computational Harmonic Analysis*, 37(3):531–544, November 2014.

[185] B. Wen, X. Chen, and T. K. Pong. Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *SIAM Journal on Optimization*, 27(1):124–145, 2017.

[186] S. J. Wright. Identifiable Surfaces in Constrained Optimization. *SIAM J. Control Optim.*, 31(4):1063–1079, July 1993.

[187] Z. Wu, C. Li, M. Li, and A. Lim. Inertial proximal gradient methods with bregman regularization for a class of nonconvex optimization problems. *Journal of Global Optimization*, 79(3):617–644, 2021.

[188] Y. Xia and Z. Xu. The performance of the amplitude-based model for complex phase retrieval. *Information and Inference: A Journal of the IMA*, 13(1), 01 2024.

[189] Z. Yang, L. F. Yang, E. X. Fang, T. Zhao, Z. Wang, and M. Neykov. Misspecified nonconvex statistical optimization for sparse phase retrieval. *Mathematical Programming*, 176:545–571, July 2019.

[190] M. B. Yu. and L. G. Sodin. On the ambiguity of the image reconstruction problem. *Optics Communications*, 30(3):304–308, 1979.

[191] Z. Yuan, H. Wang, and Q. Wang. Phase retrieval via Sparse Wirtinger Flow. *Journal of Computational and Applied Mathematics*, 355:162–173, August 2019.

[192] H. Zhang, Y. Liang, and Y. Chi. A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms. *Journal of Machine Learning Research*, 18(141):1–35, 2017.