Towards optimal algorithms for the recovery of low-dimensional models with linear rates

Yann Traonmilin, Jean François Aujol, Antoine Guennec

(2024) **00**, 1–36
https://doi.org/DOI HERE

Towards optimal algorithms for the recovery of low-dimensional models with linear rates

YANN TRAONMILIN* AND JEAN-FRANÇOIS AUJOL AND ANTOINE GUENNEC
Univ. Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251, F-33400, Talence, France
*Corresponding author: yann.traonmilin@math.u-bordeaux.fr

[Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year]

We consider the problem of recovering elements of a low-dimensional model from linear measurements. From signal and image processing to inverse problems in data science, this question has been at the center of many applications. Lately, with the success of models and methods relying on deep neural networks leading to non-convex formulations, traditional convex variational approaches have shown their limits. Furthermore, the multiplication of algorithms and recovery results makes identifying the best methods a complex task. In this article, we study recovery with a class of widely used algorithms without considering any underlying functional. This result leads to a class of projected gradient descent algorithms that recover a given low-dimensional with linear rates. The obtained rates decouple the impact of the quality of the measurements with respect to the model from its intrinsic complexity. As a consequence, we can directly measure the performance of this class of projected gradient descents through a restricted Lipschitz constant of the projection. By optimizing this constant, we define optimal algorithms. Our general approach provides an optimality result in the case of sparse recovery. Moreover, we uncover underlying linear rates of convergence for some "plug and play" imaging methods relying on deep priors by interpreting our results in this context, thus linking low-dimensional recovery and recovery with deep priors under a unified theory, validated by experiments on synthetic and real data.

Keywords: optimal algorithms; low-dimensional recovery; projected gradient descent; plug-and-play methods

1. Introduction

In this paper, we consider the general noiseless observation model in finite dimension:

$$y = A\hat{x} \tag{1}$$

where y is an m-dimensional vector of measurements, A is an under-determined linear operator from $\mathbb{R}^N$ to $\mathbb{C}^m$ (i.e. $m < n$) and $\hat{x} \in \mathbb{R}^N$ is the unknown vector we want to recover. The problem of recovering $\hat{x}$ from y is ill-posed. It is thus necessary to use prior information on $\hat{x}$ to hope for an estimation of $\hat{x}$ with a guarantee of success.

In this work, we suppose that $\hat{x}$ belongs to a (potentially infinite) union of subspaces Σ (i.e. a homogeneous set: for every $x \in \Sigma$ and $\lambda \in \mathbb{R}$, $\lambda x \in \Sigma$) that models known properties of the unknown. Examples of such models include sparse as well as low-rank models and many of their generalizations (see [14] for an overview). The problem of recovering elements of a low-dimensional model from their measurements has been at the center of inverse problems in data science. This is for instance the case for many problems in signal and image processing where the unknown $\hat{x}$ is the signal or image of interest, and the matrix A models a degradation such as a subsampling or a blur. Note that, in this article, we consider the *noiseless* case as we focus on the geometry of the estimation of elements of Σ in relation to the chosen algorithm and the measurement operator A.

To recover $\hat{x}$, a classical method is to solve the minimization problem

$$x^\star \in \arg\min \frac{1}{2}\|Ax-y\|_2^2 + \lambda R(x) \tag{2}$$

where R is a function – the regularizer – that enforces some structure on the solution $x^\star$. Without any assumption on the easiness of the calculation of $x^\star$, functions R of the form $d(\cdot,\Sigma)$ (distance to the model set for a given norm) achieve the best identifiability guarantees [7]. Unfortunately, for classical low-dimensional models such as sparse models such minimization is NP hard [20].

Another possibility is to use a convex proxy of minimization (2) (i.e. using a convex R) that guarantees the recovery of elements of Σ. A general method for the best possible choice of convex regularization R has been presented in [37]. However, in terms of practical recovery guarantees, one must combine guarantees of success of (2) with convergence guarantees of the chosen algorithm.

While the convergence of many algorithms is verified for any instance of the convex problems, for the non-convex methods, heuristics approximating the minimization of (2) are studied directly, and identifiability guarantees of the chosen algorithm are proven under conditions much more stringent than the guarantees of the ideal non-convex minimization. Moreover, it is often hard to compare results accurately with convex methods, resulting in potentially different behaviors between theoretical results and practical implementations of said methods.

In this landscape of methods for solving inverse problems, algorithms based on deep priors have become the state-of-the-art reference in domains such as image processing. These methods have in common the fact that the prior used, one way or another, is learned on a large database with a deep neural network (DNN). Such priors are called deep priors. A large part of the resulting algorithms, such as plug-and-play methods (PnP) [39] fall within a non-convex framework, which leads to local convergence and sublinear rates. Moreover, while the role of the low-dimensional model is studied in some works on deep priors, it is not extensively studied in the plug-and-play literature.

In this paper, we propose to study directly the identifiability performance of a class of recovery algorithms in order to answer the question:

Given a low-dimensional model Σ, what is the optimal algorithm (in a given specific class of algorithms) to recover elements of Σ from linear observations?

We propose to study the problem of low-dimensional recovery beyond variational methods where a functional (to be minimized) must be explicit (in particular a regularizer). The objective is to obtain sharp guarantees and to facilitate comparisons between different algorithms (originated from convex or non-convex methods). Once guarantees are obtained, we can define the optimal method as the one having the best guarantees. Moreover, the framework allows us to consider algorithms relying on deep priors and to understand their identifiability and convergence properties. To achieve this objective, we first need to specify the chosen class of algorithms and then what notion of optimality we use.

1.1. *Method of averaged directions*

Of course, many choices of classes of methods are possible. In this article, we focus on iterative algorithms of the form

$$x_{n+1} = x_n - \mu d_n \tag{3}$$

where we decompose

$$d_n = A^H(Ax_n - y) + g_n \tag{4}$$

where $g_n = g(x_n)$ is a function of x_n only (the first iterate is calculated from an initial value $x_0 \in \mathbb{R}^N$). At each step, a direction is calculated as the average of a data-fit direction and a regularizing direction. Such algorithms can be viewed as the result of a design without the help of an underlying functional to minimize: the iterations are built by averaging

- a back-projection of the residual between the measurements y and the current iterate;
- and a direction g_n pushing towards the low-dimensional model Σ (that might not be the gradient of a regularization function).

We can model both convex and "non-convex" approaches for sparse recovery and beyond.

- In the convex case, we can set $g_n = \lambda \nabla R(x_n)$. We fall exactly on the gradient descent for the optimization of (2).
- In the non-convex case, we can chose $g_n = (I - A^H A)(x_n - P_\Sigma(x_n))$ where P_Σ is a projection on the model set. This algorithm is indeed a variation of the projected gradient descent (PGD) with the projection P_Σ where the projection and descent steps are exchanged (see Remark 2.1; we will keep the name projected gradient descent as a slight abuse of definition, as only this variation of PGD will be used in this article). In the case of sparse recovery, with P_Σ the hard thresholding operator, this algorithm is iterative hard thresholding [6] (or more generally iterative thresholding pursuit [13]).

  Note that this algorithm can be used to perform recovery of inverse problems with deep priors (i.e. non-convex regularization parametrized by deep neural networks (DNN) using plug-and-play methods), where the projection P_Σ is performed using a general purpose denoiser parametrized with a deep neural network [41].

1.2. *A notion of optimal algorithm for low-dimensional recovery*

In [37], a notion of optimal (convex) regularization for low-dimensional recovery was proposed. Many possibilities, depending on the desired objective, exist for the definition of optimality. In this article, we will propose an optimality condition that provides a near-optimal trade-off between the identifiability of elements of Σ and the rate of convergence for newly given state of the art recovery guarantees.

In the case where elements $\hat{x} \in \Sigma$ are uniquely identifiable from y, we consider algorithms where we can guarantee *uniform recovery* with a global convergence, i.e. given A (with the adequate properties) and any $\hat{x}$:

$$\forall \hat{x} \in \Sigma, \|x_n - \hat{x}\|_2 \to_{n\to+\infty} 0 \tag{5}$$

for any choice of initialization x_0. Note that if any $\hat{x} \in \Sigma$ is identifiable from $A\hat{x}$, then the operator A necessarily has a lower restricted isometry property (RIP) [7] (a property that will be central in our analysis).

Within the chosen set of algorithms, we will look for algorithms achieving linear rates of convergence, under a RIP assumption on A, i.e. there is $r \in [0;1)$ such that, for all $n \geq 0$,

$$\|x_{n+1} - \hat{x}\|_2^2 \leq r\|x_n - \hat{x}\|_2^2. \tag{6}$$

In other words, we specifically search for algorithms that perform fast recovery of the elements of the model (as was demonstrated for iterative thresholding algorithms for sparse recovery with some

conditions on A [6]). Specifically, we ask ourselves what choice of directions d_n (i.e. choice of g_n) yields the best rates and recovery guarantees. Note that general convex methods might yield sublinear rates of convergence. Our study thus excludes such cases. We also do not add any constraint onto the computational complexity of each iteration. However, the class of average directions that we chose naturally yields the class of projected gradient descent algorithms. In this case, we do not take into account the complexity of the computation of the projection P_Σ. Although it is often fast to compute in the case of sparse modeling (in that case, projection is just a thresholding), there are other instances when it can be much more demanding (i.e. projection on sets of low-rank tensors). For plug-and-play methods with deep priors, the projection is realized by a forward pass in a deep neural network, which is a fast operation.

1.3. *Contributions*

In this article, we propose a framework aiming towards the study of the optimality of low-dimensional recovery algorithms for a specific class of averaged directions.

- In Section 2, we propose a general recovery result for the methods of averaged direction where we specifically choose the data-fit direction as a gradient of the ℓ^2-function (we can also interpret this choice purely geometrically without an underlying function). Our analysis leads to a natural choice of regularizing direction, which gives a projected gradient descent with a given projection P_Σ onto Σ. The obtained linear rate decouples the quality of the measurements through the restricted isometry constant of A and the quality of the chosen algorithm through a newly introduced *restricted* Lipschitz condition on P_Σ. This enables the introduction of a precise optimality notion for the class of projected gradient descent algorithms relying on this *restricted* Lipschitz condition.
- In Section 3, we investigate optimal projections optimizing the *restricted* Lipschitz condition. We show that the orthogonal projection $P_\Sigma^\perp$ onto a general union of subspaces (when it exists) always satisfies this condition, and that, in the context of sparse recovery, it is indeed optimal when considering the family of model sets Σ_k for all levels of sparsity k. This result also shows that for fixed sparsity, while the optimal projection may not be the orthogonal projection, its *restricted* Lipschitz constant is close to the constant of the orthogonal projection. Also note that finding projections with optimal *restricted* Lipschitz constants is of mathematical interest independently of the context of low-dimensional recovery.
- In Section 4, we interpret our results in the context of plug-and-play methods for solving inverse problems with deep priors. We show that, when we explicit the underlying low-dimensional model and the corresponding projection that is performed, the so-called proximal gradient descent plug-and-play method (that can be interpreted as a projected gradient descent) exhibits linear rates of convergence towards elements of the model induced by the chosen denoiser if the *restricted*-Lipschitz property is verified. In the context of PnP, convergence in this non-convex setting is generally shown to be sub-linear to the minimum of a given function. Moreover, the role of the low-dimensional model in such algorithms has not been often explored. We show experimentally this global linear convergence suggesting that the *restricted*-Lipschitz property is indeed verified approximately for some general purpose denoisers (DRUNet denoisers in our case).

1.4. *Related work*

Our work aims at building a framework for the design of optimal algorithms for the recovery of low-dimensional models. This follows some ideas from [37] where optimal convex regularizers are defined

and calculated. In [25], the authors propose a similar framework to design a possibly non-convex regularizer from a data source. Another direction of research related to optimal methods is studying the intrinsic computational complexity of recovery algorithms in the context of inverse problems [4, 8, 11, 32]. Our approach mainly relies on restricting the class of considered algorithms to be able to find non-trivial optimal algorithms.

The projected gradient descent, which is at the center of this paper has been widely studied in many applications. For sparse recovery, the projected gradient descent is called iterative hard thresholding (or more generally hard thresholding pursuit) and has been shown to linearly converge to the unknown under a restricted isometry property of the observation operator A [6, 13]. For the recovery of low-dimensional models, conditions for the linear convergence of projected gradient descent with approximate projections are given in [15]. In [2], global convergence of PGD is given for a class of generalized sparsity models and the orthogonal projection. Thanks to our convergence analysis, our work explicitly links the convergence rate of PGD with a restricted Lipschitz constant of the considered (general) projection. Within thresholding algorithms, [27] studies optimal thresholding operators with respect to a local concavity property used to give local convergence properties of projected gradient descent for minimizing general functions in [3]. Another local convergence analysis is provided in [40]. In our work, we only consider global convergence. In [24], it is proposed to learn an optimal non-linearity with a data-driven method. Note that global convergence of gradient projection has been shown under a general KL property in [1]. General stationary properties of the iterates of PGD are given in [28]. In this work, we focus on linear rates of convergence to solutions of linear inverse problems.

One objective of this paper is to give insights into the geometry of algorithms using deep priors to solve inverse problems, mainly in imaging. The recent literature on the subject is profuse (see [29, 33] for an overview). In particular, many variations of plug-and-play methods (where a deep neural network is used to perform an iteration of an optimization algorithm) exist, each one corresponding to a variation of an optimization algorithm such as forward-backward, ADMM, etc...[9, 10, 23, 39]. Specific designs of the regularizing direction have been given to guarantee the sublinear convergence of such methods under global Lipschitz condition [19] and differentiability hypothesis of the functional to minimize [21]. The control of the global Lipschitz constant of the DNN architecture is a also at the center of other works [38]. Our work shows that the control of a weaker restricted Lipschitz constant is sufficient for low-dimensional recovery. We must also cite methods where the projection onto the model set Σ is explicitly using, e.g auto-encoders [30] or other generative models such as variational auto-encoders [16]. Recovery guarantees of low-dimensional models are studied under very stringent conditions (Gaussian measurements) in [18]. In [34], a general study of the identifiability of models obtained in a learning context is performed.

1.5. *Notations, definitions and preliminaries*

Given a linear operator $A : \mathbb{R}^N \to \mathbb{C}^m$, we denote A^H its Hermitian adjoint. Given a function $f : \mathbb{R}^N \to \mathbb{R}^N$, we call $\mathrm{Fix}(f) = \{x \in \mathbb{R}^N : f(x) = x\}$ the set of fixed points of f.

We use the following definition of restricted isometry constant.

Definition 1.1 *The operator A has restricted isometry constant $\delta < 1$ on the secant set $\Sigma - \Sigma = \{x_1 - x_2 : x_1, x \in \Sigma\}$ if for all $x_1, x_2 \in \Sigma$*

$$\|(A^H A - I)(x_1 - x_2)\|_2 \leq \delta \|x_1 - x_2\|_2 \tag{7}$$

We write $\delta_\Sigma(A)$ the smallest admissible restricted isometry constant (RIC).

The existence of the RIC $\delta_\Sigma(A) < 1$ implies a restricted isometry property (RIP) of the operator A in the traditional sense defined by Equation (8) [14] as shown in the following Lemma.

Lemma 1.1 *Suppose $\delta_\Sigma(A) < 1$. Let $\Sigma \subset \mathbb{R}^N$ and $x_1, x_2 \in \Sigma$. We have*

$$(1-\delta_\Sigma(A))\|x_1-x_2\|_2^2 \leq \|A(x_1-x_2)\|_2^2 \leq (1+\delta_\Sigma(A))\|x_1-x_2\|_2^2. \tag{8}$$

Proof We have, for $x_1, x_2 \in \Sigma$, $x_1 \neq x_2$,

$$\begin{aligned}\left|\frac{\|A(x_1-x_2)\|_2^2-\|x_1-x_2\|_2^2}{\|x_1-x_2\|_2^2}\right| &= \left|\frac{\langle A(x_1-x_2),A(x_1-x_2)\rangle-\langle x_1-x_2,x_1-x_2\rangle}{\|x_1-x_2\|_2^2}\right| \\ &= \left|\frac{\langle (A^HA-I)(x_1-x_2),x_1-x_2\rangle}{\|x_1-x_2\|_2^2}\right|\end{aligned} \tag{9}$$

Using the Cauchy-Schwarz inequality we have

$$\begin{aligned}\left|\frac{|\|A(x_1-x_2)\|_2^2-\|x_1-x_2\|_2^2}{\|x_1-x_2\|_2^2}\right| &\leq \left|\frac{\|(A^HA-I)(x_1-x_2)\|_2\|x_1-x_2\|_2}{\|x_1-x_2\|_2^2}\right| \\ &= \frac{\|(A^HA-I)(x_1-x_2)\|_2}{\|x_1-x_2\|_2}\end{aligned} \tag{10}$$

With the definition of RIC,

$$\left|\frac{|\|A(x_1-x_2)\|_2^2-\|x_1-x_2\|_2^2}{\|x_1-x_2\|_2^2}\right| \leq \delta_\Sigma(A). \tag{11}$$

□

To simplify calculations, in the following we will consider operators A that have been centered to have the best possible RIC δ. This eliminates the typical problem with the RIP hypothesis that multiplying the measurement operator by a factor does not change recovery capabilities but can worsen the RIC.

Definition 1.2 *We say that A is centered for the RIC if for all $\lambda \in \mathbb{R}$, $\delta_\Sigma(A) \leq \delta_\Sigma(\lambda A)$.*

Hence, if A is not centered we consider instead $\tilde{A} = \lambda_0 A$ such that $\delta_\Sigma(\tilde{A}) = \inf_{\lambda \geq 0} \delta_\Sigma(\lambda A)$.

Note that the lower RIP is a necessary condition for the identifiability of all elements of Σ from measurements with the operator A [7]. Consequently, in the context of uniform recovery of Σ from linear measurements in finite dimension, we can always make a RIP hypothesis, otherwise such uniform recovery is not possible.

We define projections and orthogonal projections as they will take a central role in this article.

Definition 1.3 (Projection) *Let $\Sigma \subset \mathbb{R}^N$. A (set-valued) projection onto Σ is a function P such that for any $z \in \mathbb{R}^N$, $P(z) \subset \Sigma$.*

By abuse of notation, to facilitate reading, an equation true for any $w \in P(z)$ is written using the notation $P(z)$.

Definition 1.4 (Orthogonal projection) *We define, when it exists, the (set-valued) orthogonal projection onto a set $\Sigma \subset \mathbb{R}^N$ as follows: for all $z \in \mathbb{R}^N$*

$$P_\Sigma^\perp(z) = \arg\min_{x \in \Sigma} \|x - z\|_2^2. \tag{12}$$

2. On optimal low-dimensional recovery with averaged directions under a linear convergence rate

In this section, we first discuss our specific choice of data-fit directions h_n and regularizing direction g_n in the iterations

$$x_{n+1} = x_n - \mu(h_n + g_n) \tag{13}$$

initialized at an arbitrary $x_0 \in \mathbb{R}^N$. We then give our main theorem guaranteeing recovery for a specific subclass of regularizing directions which appears naturally in our proof. This general recovery result is the basis for our definition of optimal recovery.

2.1. *Design of the data-fit direction h_n*

To design an iterative algorithm, it is natural to first choose a direction that pushes towards $\hat{x}$ from x_n, i.e. the best approximation of the direction $x_n - \hat{x}$. As we have access to $r_n = Ax_n - y = A(x_n - \hat{x})$, the most common direction that helps convergence to an estimation of $\hat{x}$ is the back-projection of the residual

$$h_n = A^H A(x_n - \hat{x}), \tag{14}$$

which matches the gradient of an ℓ^2 data-fit functional. We call h_n the data-fit direction. If x_n is close to Σ, under a restricted isometry hypothesis (Definition 1.1), we have that $A^H r_n = A^H A(x_n - \hat{x}) \approx x_n - \hat{x}$. Also, if $g_n = 0$ we fall on the Landweber iteration: $d_n = h_n = \nabla F(x_n)$ with $F(x) = \frac{1}{2}\|Ax_n - y\|_2^2$ (gradient descent of under-determined least-squares functional). In the following, we use this specific choice of h_n (which can be interpreted in a purely geometrical way without considering F). Other possibilities could be the (sub)-gradient of other data-fit functionals such as the ℓ^1 data-fit for robust regression. We can also mention implicit schemes such as the forward-backward algorithm where the gradient direction is estimated at x_{n+1} instead of x_n. Such generalizations are left for future work as the choice of a data-fit direction is also linked with the type of noise in the noisy case (which is out of the scope of this paper).

2.2. *Design of the regularizing direction g_n*

Now that we have chosen a data-fit direction, given Σ what is the best way to make x_n converge towards an element of Σ? Using the direction $x_n - P_\Sigma^\perp(x_n)$ where $P_\Sigma^\perp$ is an orthogonal projection onto Σ (that we suppose existing) seems the best as it is the shortest path between x_n and Σ for the ℓ^2 metric. As enforcing the model Σ is a central point in low-dimensional recovery, we can also consider the class of projected descent algorithms (see Remark 2.1)

$$x_{n+1} = P_\Sigma(x_n - \mu A^H A(x_n - \hat{x})) \tag{15}$$

where P_Σ is a projection onto Σ. For sparse recovery, these algorithms include iterative hard thresholding (or hard thresholding pursuit) and iterative soft thresholding which is just the subgradient descent

of the ℓ^1 regularization with ℓ^2 data-fit. We can always rewrite these algorithms as a method of averaged direction. With the corresponding regularizing direction $g_n = -\frac{1}{\mu}(x_{n+1} - x_n) - h_n$, we have the following equation

$$x_{n+1} = x_n - \mu(A^H A(x_n - \hat{x}) + g_n). \tag{16}$$

Our objective is to give a framework aiming towards the calculation of optimal regularizing directions g_n. Note that, if we want this question to have an interesting meaning from an optimization point of view, we should consider regularizing directions as functions of x_n that do not depend on $y = A\hat{x}$. Indeed, we could just choose g_n depending on $\arg\min_{x\in\Sigma} \|Ax - y\|_2^2$ for a convergence in one iteration. We will see that with this restriction on g_n, we naturally fall on projected descent algorithms. For this class of algorithms, we give linear rates of convergence, and we show in Section 3, in the context of sparse recovery that the orthogonal projection is optimal (when considering the whole collection of sparse models) for these rates.

2.3. *Recovery of low-dimensional models with linear rates*

In this Section, we show that the specific class of averaged directions defining a projected gradient descent can recover low-dimensional models provided the projection has a *restricted* Lipschitz condition.

We first give an equivalent condition on g_n for the linear recovery of $\hat{x}$ (defined in (6)). We consider a uniform linear rate for all iterations, i.e. no finite time "burning" period is allowed in our study (see Section 5).

Lemma 2.1 (Characterization of linear convergence) *Let $\hat{x} \in \mathbb{R}^N, A \in \mathbb{C}^{m\times N}$. Consider the iterations*

$$x_{n+1} = x_n - \mu(A^H A(x_n - \hat{x}) + g_n). \tag{17}$$

Then $x_n \to_{n\to\infty} \hat{x}$ at a uniform linear rate $r < 1$ (i.e. (6) is verified for all $n \geq 0$) if and only if for all $n \geq 0$, we have

$$(1-r)\|e_n\|_2^2 + \mu^2\|A^H A e_n + g_n\|_2^2 - 2\mu\langle e_n, g_n\rangle \leq 2\mu\|Ae_n\|_2^2 \tag{18}$$

where $e_n = x_n - \hat{x}$.

Proof Recall that linear recovery with uniform rate $r < 1$ (defined in (6)) is equivalent to: for all $n \geq 0$,

$$\|x_{n+1} - \hat{x}\|_2^2 \leq r\|x_n - \hat{x}\|_2^2. \tag{19}$$

As we have, with $d_n := A^H A(x_n - \hat{x}) + g_n$,

$$\begin{aligned} \|x_{n+1} - \hat{x}\|_2^2 &= \|x_n - \mu d_n - \hat{x}\|_2^2 \\ &= \|x_n - \hat{x}\|_2^2 - 2\mu\langle x_n - \hat{x}, d_n\rangle + \mu^2\|d_n\|_2^2. \end{aligned} \tag{20}$$

Hence linear convergence rate with uniform rate r is equivalent to showing that for all $n \geq 0$,

$$\|x_n - \hat{x}\|_2^2 - 2\mu\langle x_n - \hat{x}, d_n\rangle + \mu^2\|d_n\|_2^2 \leq r\|x_n - \hat{x}\|_2^2 \tag{21}$$

which is equivalent to

$$(1-r)\|x_n-\hat{x}\|_2^2+\mu^2\|d_n\|_2^2 \leq 2\mu\langle x_n-\hat{x},d_n\rangle. \tag{22}$$

Let $e_n = x_n-\hat{x}$. We have $d_n = A^HAe_n+g_n$. Hence (22) is equivalent to the following inequalities:

$$\begin{aligned}(1-r)\|e_n\|_2^2+\mu^2\|A^HAe_n+g_n\|_2^2 &\leq 2\mu\langle e_n,A^HAe_n+g_n\rangle\\ (1-r)\|e_n\|_2^2+\mu^2\|A^HAe_n+g_n\|_2^2 &\leq 2\mu\langle Ae_n,Ae_n\rangle+2\mu\langle e_n,g_n\rangle\\ (1-r)\|e_n\|_2^2+\mu^2\|A^HAe_n+g_n\|_2^2 &\leq 2\mu\|Ae_n\|_2^2+2\mu\langle e_n,g_n\rangle.\end{aligned} \tag{23}$$

We finally obtain the equivalent condition

$$(1-r)\|e_n\|_2^2+\mu^2\|A^HAe_n+g_n\|_2^2-2\mu\langle e_n,g_n\rangle \leq 2\mu\|Ae_n\|_2^2. \tag{24}$$

□

This Lemma highlights that for a given iteration n, to get the smallest value of r (i.e. the fastest convergence), the best possible choice of direction g_n is to minimize the left-hand side of (18), i.e. we would have to minimize

$$H(g) := \mu^2\|A^HAe_n+g\|_2^2-2\mu\langle e_n,g\rangle \tag{25}$$

with respect to the direction g. However, without adding any constraint on g, this would require knowledge of the unknown $\hat{x}$.

Instead, we provide regularizing directions g_n that guarantee linear convergence with a bound on H, and that we conjecture to be near-optimal (see Remark 2.3). We will see that these directions g_n correspond to variants of projected descent, i.e iterations of the form

$$\begin{aligned}z_n &= P_\Sigma(x_n)\\ x_{n+1} &= z_n+\mu A^H(Az_n-y)\end{aligned} \tag{26}$$

where P_Σ is a projection onto Σ, i.e. a function $\mathbb{R}^N \to \Sigma$. The condition for recovery with a linear rate is a restricted Lipschitz condition of the projection (weaker than the classical Lipschitz condition).

Definition 2.1 (Restricted Lipschitz property) *Let $P:\mathbb{R}^N\to\mathbb{R}^N$. Then P has the restricted β-Lipschitz property with respect to Σ if for all $z\in\mathbb{R}^N, x\in\Sigma$ we have*

$$\|P(z)-x\|_2 \leq \beta\|z-x\|_2 \tag{27}$$

Note that any P verifying this condition is such that $\Sigma\subset \mathrm{Fix}(P)$ as $\|z-x\|_2=0$ (i.e. $z=x$) implies $P(z)=x$ (see also Lemma 3.1). Hence, this is a weaker statement than the classical Lipschitz property $\|P(z)-P(x)\|_2\leq\beta\|z-x\|_2$, since one variable is restricted to Σ.

We note $\beta_\Sigma(P)$ the smallest β such that P has the restricted β-Lipschitz property.

For set-valued projections we say that P has the restricted β-Lipschitz property if for all $z \in \mathbb{R}^N, x \in \Sigma, y \in P(z)$, $\|y-x\|_2 \leq \beta \|z-x\|_2$.

We expect that a restricted Lipschitz constant is such that $\beta \geq 1$ (with equality for the orthogonal projection onto a linear subspace, see Lemma 3.4). In Section 3, we show in a general setting the existence of projections with the restricted β-Lipschitz property (in particular, orthogonal projections onto Σ).

We can now give a general result that guarantees recovery with a linear rate under the RIP and this restricted Lipschitz condition.

Theorem 2.1 (Linear recovery of low-dimensional models) *Let $\Sigma \subset \mathbb{R}^N$. Suppose A is centered for the RIC with RIC $\delta = \delta_\Sigma(A)$. Suppose $\mu = 1$. For any $\hat{x} \in \Sigma, x_0 \in \mathbb{R}^N$, consider the iterates x_n resulting from Algorithm (13) with*

- $h_n = A^H A(x_n - \hat{x})$;
- $g_n = (I - A^H A)(x_n - P_\Sigma(x_n))$ *where P_Σ is a projection onto Σ.*

Suppose P_Σ has the β-restricted Lipschitz property with respect to Σ (with $\beta = \beta_\Sigma(P_\Sigma)$). Suppose $\delta^2\beta^2 < 1$, then we have

1. ***Linear convergence rate**: for all $\hat{x} \in \Sigma, x_0 \in \mathbb{R}^N, n \geq 1$,*

$$\|x_n - \hat{x}\|_2^2 \leq (\delta^2\beta^2)^n \|x_0 - \hat{x}\|_2^2. \tag{28}$$

2. ***Necessary Lipschtiz condition**: the restricted β-Lipschitz condition of P_Σ is necessary to obtain such a uniform linear rate of convergence.*

Proof **Proof of property 1.** We suppose $\mu = 1$ and we prove Equation (28). Let $e_n = x_n - \hat{x}$. To show recovery, we will bound H defined in (25) and use Lemma 2.1. To put our proof in a broader context, we first consider a generic direction g that would maximize the speed of convergence (minimizing r), i.e. minimize the function H. This will justify our specific choice of directions g_n. We rewrite H as a least-squares expression in g. We have

$$\begin{aligned} H(g) - \|A^H A e_n\|_2^2 &= \|A^H A e_n + g\|_2^2 - 2\langle e_n, g\rangle - \|A^H A e_n\|_2^2 \\ &= \|g\|_2^2 + 2\langle A^H A e_n, g\rangle - 2\langle e_n, g\rangle \\ &= \|g\|_2^2 + 2\langle (A^H A - I)e_n, g\rangle + \|(A^H A - I)e_n\|_2^2 - \|(A^H A - I)e_n\|_2^2 \end{aligned} \tag{29}$$

Hence,

$$F(g) := \|g + (A^H A - I)e_n\|_2^2 = H(g) - \|A^H A e_n\|_2^2 + \|(A^H A - I)e_n\|_2^2. \tag{30}$$

We deduce

$$\arg\min_g F(g) = \arg\min_g H(g). \tag{31}$$

In the following, instead of explicitly minimizing F (and H), we provide a bound that guarantees convergence for our choice of regularizing direction.

As we consider the case when g must not depend on the current residual and is only a function of x_n enforcing some prior on the solution: i.e. $g = f(x_n)$, i.e.

$$F(g) = F(f(x_n)) = \|f(x_n) - (I - A^H A)e_n\|_2^2, \tag{32}$$

it is natural to introduce a direction towards Σ (see Section 2.2). This can be done using an arbitrary projection P_Σ on Σ. Let $w_n = P_\Sigma(x_n) - \hat{x} = x_n - \hat{x} - (x_n - P_\Sigma(x_n)) = e_n - (x_n - P_\Sigma(x_n))$, we have

$$F(g) = \|f(x_n) - (I - A^H A)w_n - (I - A^H A)(x_n - P_\Sigma(x_n))\|_2^2. \tag{33}$$

For the term $(I - A^H A)w_n$, with the RIC (Definition 1.1), we have

$$\|(I - A^H A)w_n\|_2^2 \leq \delta^2 \|w_n\|_2^2. \tag{34}$$

Moreover, we supposed that there exist elements $u_1 - u_2 = w_n \in \Sigma - \Sigma$ making the above inequality tight as the RIC is chosen as small as possible. Hence for any projection P_Σ such that $P_\Sigma(\mathbb{R}^N) = \Sigma$, this inequality is possibly reached (by setting $P_\Sigma(x_0) = u_1, \hat{x} = u_2$). Hence choosing $g = g_n = (I - A^H A)(x_n - P_\Sigma(x_n))$, guarantees

$$F(g) = \|(I - A^H A)w_n\|_2^2 \leq \delta^2 \|w_n\|_2^2 \tag{35}$$

and that this inequality is tight in the sense that it is reached for some initialization x_0 and unknown $\hat{x}$.

With this choice of g_n, we have

$$d_n = h_n + g_n = A^H(Ax_n - y) + (I - A^H A)(x_n - P_\Sigma(x_n)). \tag{36}$$

Going back to H with Equations (30) and (35),

$$\begin{aligned} H(g_n) &= F(g_n) + \|A^H A e_n\|_2^2 - \|(A^H A - I)e_n\|_2^2 \\ &\leq \delta^2 \|w_n\|_2^2 + \|A^H A e_n\|_2^2 - \|(A^H A - I)e_n\|_2^2. \end{aligned} \tag{37}$$

We deduce the following bound on the left side of condition (18):

$$\begin{aligned} &(1-r)\|e_n\|_2^2 + \|A^H A e_n + g_n\|_2^2 - 2\langle e_n, g_n \rangle \\ &= (1-r)\|e_n\|_2^2 + H(g_n) \\ &\leq (1-r)\|e_n\|_2^2 + \delta^2 \|w_n\|_2^2 + \|A^H A e_n\|_2^2 - \|(A^H A - I)e_n\|_2^2 \\ &= (1-r)\|e_n\|_2^2 + \delta^2 \|w_n\|_2^2 + \|A^H A e_n\|_2^2 - \|A^H A e_n\|_2^2 - \|e_n\|_2^2 + 2\langle Ae_n, Ae_n \rangle \\ &= -r\|e_n\|_2^2 + \delta^2 \|w_n\|_2^2 + 2\|Ae_n\|_2^2. \end{aligned} \tag{38}$$

Using the restricted β-Lipschitz property of P_Σ, we have $\|w_n\|_2^2 = \|P_\Sigma(x_n) - \hat{x}\|_2^2 \leq \beta^2\|x_n - \hat{x}\|_2^2 = \beta^2\|e_n\|_2^2$ and, with $r = \delta^2\beta^2$,

$$\begin{aligned}(1-r)\|e_n\|_2^2 + \|A^HAe_n + g_n\|_2^2 - 2\langle e_n, g_n\rangle &\leq (\delta^2\beta^2 - r)\|e_n\|_2^2 + 2\|Ae_n\|_2^2 \\ &= 2\|Ae_n\|_2^2.\end{aligned} \tag{39}$$

which is exactly condition (18).

Proof of property 2.

The necessity of the restricted Lipschitz condition comes from the following fact. If $\beta_\Sigma(P_\Sigma) = +\infty$, we can find $x_0, \hat{x}$ such that $\frac{\|P_\Sigma(x_0)-\hat{x}\|_2}{\|x_0-\hat{x}\|_2} \to \infty$. With the RIP on A, this implies that $\frac{\|x_1-\hat{x}\|_2}{\|x_0-\hat{x}\|_2} \to \infty$. □

Remark 2.1 *Note that the choice of g_n in this Theorem is exactly the projected gradient descent where the projection and descent steps are inverted (i.e. iterative hard thresholding or hard thresholding pursuit for sparse recovery and orthogonal projection). Indeed, we have*

$$\begin{aligned}x_{n+1} &= x_n - \left(A^HA(x_n - \hat{x}) + (I - A^HA)(x_n - P_\Sigma(x_n))\right) \\ &= x_n - A^HAx_n + A^HA\hat{x} - x_n + P_\Sigma(x_n) + A^HA(x_n - P_\Sigma(x_n)) \\ &= P_\Sigma(x_n) - A^HA(P_\Sigma(x_n) - \hat{x}).\end{aligned} \tag{40}$$

This theorem tells us that projected gradient descent (PGD) guarantees the recovery of low-dimensional with linear rates under a RIP condition provided a restricted Lipschitz property of the projection. Optimal PGD for those guarantees are those whose projection minimizes the restricted Lipschitz constant; it quantifies two properties of the algorithm:

- the identifiability properties of PGD: if $\delta_\Sigma(A) < \frac{1}{\beta}$, then $x_n \to \hat{x}$;
- the rate of convergence: for a fixed A, the smaller the restricted Lipschitz constant $\beta_\Sigma(P_\Sigma)$, the faster the recovery of $\hat{x}$.

Given these two facts, we propose a quantitative optimality measure (in terms of convergence) of a projected descent algorithm parametrized by a projection P with $\beta_\Sigma(P)$.

Definition 2.2 (Optimal projection) *We define the optimal projection $P^\star$ for the uniform recovery of a low-dimensional model set Σ with projected gradient descent with a uniform linear rate (given by Theorem 2.1) as*

$$P^\star \in \arg\min_{P \in \Pi_\Sigma} \beta_\Sigma(P) \tag{41}$$

where Π_Σ is the set of projections onto Σ having a restricted β-Lipschitz property.

In the next Section, we investigate the problem of finding optimal projections and focus on the *orthogonal projection*. We show that it is restricted Lipschitz for general unions of subspaces (i.e. $\Pi_\Sigma \neq \emptyset$). In particular, we show that the orthogonal projection is near-optimal for sparse recovery. We will see that our result can also be used to interpret the convergence of a certain class of plug-and-play algorithms for imaging inverse problems with deep priors (Section 4). For more general model sets, the question of the existence of an optimal projection is open.

Remark 2.2 *A question that naturally arises is the tightness of Theorem 2.1. We observe that the rate of convergence should be tight if the worst initialization for the restricted Lipschitz condition of the projection P_Σ matches the worst case for the RIC constant of A. While in some cases, it should be possible to construct such A, the relation between the RIC of A and the properties of P_Σ might be more intricate in general.*

Remark 2.3 *While we will focus now on optimality within projected gradient descent algorithms, we conjecture that the choice of this class of algorithms (i.e. this choice of regularizing direction) in Theorem 2.1 is close to optimal. Consider a general averaged direction algorithm with the choice $d_n = A^H A(x_n - \hat{x}) + g(x_n)$ where g is a function that does not depend on $\hat{x}$. Then*

$$\frac{\|x_{n+1} - \hat{x}\|_2}{\|x_n - \hat{x}\|_2} = \frac{\|(I - \mu A^H A)(x_n - \hat{x}) - \mu g(x_n)\|_2}{\|x_n - \hat{x}\|_2}. \tag{42}$$

The quantity $\frac{\|(I-\mu A^H A)(x_n-\hat{x})\|_2}{\|x_n-\hat{x}\|_2}$ is not bounded by a constant < 1 as $x_n \notin \Sigma$ in general. Moreover the quantity $\frac{\|g(x_n)\|_2}{\|x_n-\hat{x}\|_2}$ cannot be bounded as well (take $x_n - \hat{x} \to 0$ except for $g(x_n) = 0$). The only set where $I - \mu A^H A$ can be appropriately bounded is $\Sigma - \Sigma$ if we suppose that the RIP is a minimal assumption.

Let $\tilde{x} \in \Sigma$, we have

$$\frac{\|x_{n+1} - \hat{x}\|_2}{\|x_n - \hat{x}\|_2} = \frac{\|(I - \mu A^H A)(\tilde{x} - \hat{x}) + (I - \mu A^H A)(x_n - \tilde{x}) - \mu g(x_n)\|_2}{\|x_n - \hat{x}\|_2}. \tag{43}$$

We immediately see that setting $\mu g(n) = (I - \mu A^H A)(x_n - \tilde{x}))$, the second and third term cancel each other, leaving only to bound $\frac{\|(I-\mu A^H A)(\tilde{x}-\hat{x})\|_2}{\|x_n-\hat{x}\|_2}$ with the RIP. In other words, by hypothesis on g_n, $\tilde{x} \in \Sigma$ must depend on x_n i.e. there is a projection P_Σ such that $\tilde{x} = P_\Sigma(x_n)$. Finally, the choice $\mu = 1$ comes from the fact that we supposed $A^H A$ optimally centered for the RIC.

3. On optimal projections for projected gradient descent

In this section, we show in the general case of unions of subspaces that the orthogonal projection has a restricted Lipschitz constant. The context of unions of subspaces allows us to treat many generalized sparsity models such as group sparsity (without overlap) [5] and sparsity in levels [26] and even low-rank recovery [12]. For sparse recovery, we give an optimality result for the orthogonal projection (which corresponds to iterative hard thresholding).

3.1. *Restricted Lipschitz constant of the orthogonal projection*

We begin with the following two technical Lemmas that determine the worst case for the Lipschitz condition. The first one reformulates the expression of the Lipschitz constant.

Lemma 3.1 (Characterization of the restricted Lipschitz property) *Let $u, x \in \Sigma, z \in \mathbb{R}^N, \beta > 0$ and define*

$$Q_{\beta^2}(u, z, x) := \|u\|_2^2 - \beta^2 \|z\|_2^2 - 2\langle u - \beta^2 z, x\rangle + (1 - \beta^2)\|x\|_2^2. \tag{44}$$

Then a projection $P : \mathbb{R}^N \to \Sigma$ has restricted Lipschitz constant $\beta > 0$ (Definition 2.1) if and only if

$$\sup_{z \in \mathbb{R}^N} \sup_{x \in \Sigma} Q_{\beta^2}(P(z), z, x) \leq 0 \tag{45}$$

Proof Let $z \in \mathbb{R}^N, x \in \Sigma$. The following inequalities are equivalent:

$$\begin{aligned} \|P(z)-x\|_2^2 &\leq \beta^2 \|z-x\|_2^2 \\ \|P(z)-x\|_2^2 - \beta^2\|z-x\|_2^2 &\leq 0 \\ \|P(z)\|_2^2 - \beta^2\|z\|_2^2 - 2\langle P(z),x\rangle + 2\beta^2\langle z,x\rangle + (1-\beta^2)\|x\|_2^2 &\leq 0 \\ \|P(z)\|_2^2 - \beta^2\|z\|_2^2 - 2\langle P(z)-\beta^2 z,x\rangle + (1-\beta^2)\|x\|_2^2 &\leq 0. \end{aligned} \tag{46}$$

The last inequality is exactly $Q_{\beta^2}(P(z),z,x) \leq 0$. □

In a general setting where the orthogonal projection onto Σ exists (see Definition 1.4), we can maximize the function $Q_{\beta^2}(u,z,x)$ with respect to $x \in \Sigma$. If Σ is a finite union of subspaces it is immediate that $P_\Sigma^\perp$ exists as it is the minimum of the finite number of projections onto the individual subspaces. For sparse recovery, i.e. $\Sigma_k = \{x \in \mathbb{R}^N : \|x\|_0 \leq k\}$, $P_\Sigma^\perp$ is the hard-thresholding operator. For some other low-dimensional models such as low rank models, the union is infinite but the orthogonal projection still exists (singular value thresholding). We recall some properties of the orthogonal projection onto union of subspaces.

Lemma 3.2 *Suppose Σ is a union of subspaces and $P_\Sigma^\perp$ exists. Then for all $z \in \mathbb{R}^N$,*

1. *there exists a subspace $W \subset \Sigma$ such that $P_W^\perp(z) \in P_\Sigma^\perp(z)$;*
2. *for all linear subspaces $V \subset \Sigma$, $\|P_W^\perp(z)-z\|_2^2 \leq \|P_V^\perp(z)-z\|_2^2$;*
3. *for all linear subspaces $V \subset \Sigma$, $\|P_W^\perp(z)\|_2^2 \geq \|P_V^\perp(z)\|_2^2$;*
4. *$\langle P_W^\perp(z), P_W^\perp(z)-z\rangle = 0$.*

Proof Let $y \in P_\Sigma^\perp(z)$. Take any $W \subset \Sigma$ such that $\mathrm{span}(y) \subset W$. By definition of the orthogonal projection, $\|y-z\|_2 \leq \|w-z\|_2$ for any $w \in W \subset \Sigma$, hence $P_W^\perp(z) = y$. The other properties are direct properties of the orthogonal projection on a linear subspace. □

We give the following Lemma that explicitly calculates the maximization over $x \in \Sigma$ in condition (45).

Lemma 3.3 *Let Σ be a union of subspaces. Suppose that the orthogonal projection $P_\Sigma^\perp$ onto Σ exists. Let $z,u \in \mathbb{R}^N$. We have the following properties.*

- *If $c > 1$, let*

$$x^\star := P_\Sigma^\perp\left(\frac{u-cz}{1-c}\right). \tag{47}$$

Then

$$Q_c(u,z,x^\star) := \max_{x\in\Sigma} Q_c(u,z,x). \tag{48}$$

where Q_c is defined in Lemma 3.1. We have the following expressions of the maximum:

$$Q_c(u,z,x^\star) = \|u\|_2^2 - c\|z\|_2^2 + (c-1)\|P_\Sigma^\perp\left(\frac{u-cz}{1-c}\right)\|_2^2. \tag{49}$$

- *If $c = 1$, $Q_c(u,z,x)$ is upper bounded with respect to x only if for all $x \in \Sigma$, $\langle u - z, x\rangle = 0$. In this case,*

$$Q_c(u,z,x) = \|u\|_2^2 - \|z\|_2^2 \tag{50}$$

Proof If $c = 1$, $Q_c(u,z,x) = \|u\|_2^2 - \|z\|_2^2 - 2\langle u - z, x\rangle$ is an affine function of $x \in \Sigma$. Hence, it is upper bounded only if for all $x \in \Sigma$, $\langle u - z, x\rangle = 0$.

Now let $c > 1$. First notice that since $Q_c(u,z,x)$ is a quadratic form with respect to x with a negative leading coefficient. It is thus upper bounded and $\sup_{x\in\Sigma} Q_c(u,z,x)$ exists. By definition (Equation (44)), we have

$$\begin{aligned} Q_c(u,z,x) &= -2\langle u - cz, x\rangle + (1-c)\|x\|_2^2 + C \\ &= (1-c)\left\|x - \frac{u-cz}{1-c}\right\|_2^2 + C' \end{aligned} \tag{51}$$

where C, C' are constants that do not depend on x. Removing the constant terms, we have that maximizing Q_c is equivalent to maximizing

$$\tilde{Q}_c(x) := (1-c)\left\|x - \frac{u-cz}{1-c}\right\|_2^2. \tag{52}$$

As $1 - c < 0$, this is exactly the minimization of $\|x - \frac{u-cz}{1-c}\|_2^2$ with respect to $x \in \Sigma$ which yields (as the orthogonal projection on Σ was supposed to exist)

$$x^\star = P_\Sigma^\perp\left(\frac{u-cz}{1-c}\right). \tag{53}$$

We have, using the definition of Q_c (Equation (44)),

$$\begin{aligned} Q_c(u,z,x^\star) &= \|u\|_2^2 - c\|z\|_2^2 - 2\langle u - cz, P_\Sigma^\perp\left(\frac{u-cz}{1-c}\right)\rangle + (1-c)\|P_\Sigma^\perp\left(\frac{u-cz}{1-c}\right)\|_2^2 \\ &= \|u\|_2^2 - c\|z\|_2^2 - 2(1-c)\langle \frac{u-cz}{1-c}, P_\Sigma^\perp\left(\frac{u-cz}{1-c}\right)\rangle + (1-c)\|P_\Sigma^\perp\left(\frac{u-cz}{1-c}\right)\|_2^2. \end{aligned} \tag{54}$$

With Lemma 3.2, $\langle P_\Sigma^\perp(z), z\rangle = \|P_\Sigma^\perp(z)\|_2^2 + \langle P_\Sigma^\perp(z), z - P_\Sigma^\perp(z)\rangle = \|P_\Sigma^\perp(z)\|_2^2$ (the projection direction is orthogonal to the projection). We deduce

$$\begin{aligned} Q_c(u,z,x^\star) &= \|u\|_2^2 - c\|z\|_2^2 - 2(1-c)\|P_\Sigma^\perp\left(\frac{u-cz}{1-c}\right)\|_2^2 + (1-c)\|P_\Sigma^\perp\left(\frac{u-cz}{1-c}\right)\|_2^2 \\ &= \|u\|_2^2 - c\|z\|_2^2 + (c-1)\|P_\Sigma^\perp\left(\frac{u-cz}{1-c}\right)\|_2^2. \end{aligned} \tag{55}$$

This is exactly conclusion (49). □

The next Lemma shows in the case of linear subspaces that the best restricted Lipschitz constant $\beta = 1$ is reached by the orthogonal projection.

Lemma 3.4 *Let $\Sigma = V \subset \mathbb{R}^N$ be a linear subspace. Then $P_V^\perp$ has the restricted* 1*-Lipschitz property.*

Proof Let $z \in \mathbb{R}^N, x \in V$. We have $\langle P_V^\perp(z) - z, x\rangle = 0$ (because $P_V^\perp(z) - z \perp V$). Hence, with Lemma 3.3, we have $Q_1(P_V^\perp(z), z, x^\star) = \|P_V^\perp(z)\|_2^2 - \|z\|_2^2 = -\|P_V^\perp(z) - z\|_2^2 \leq 0$. With Lemma 3.1, $P_V^\perp$ is restricted 1-Lipschitz. □

The best possible restricted Lipschitz constant β=1 is never reached when $\Sigma \subsetneq \mathbb{R}^N$ and Span(Σ)=$\mathbb{R}^N$(e.g in the sparse or low-rank case). Hence, a constant $c > 1$ is expected in challenging cases.

Lemma 3.5 *Let $\Sigma \subsetneq \mathbb{R}^N$ be a union of subspaces such that Span$(\Sigma) = \mathbb{R}^N$. For any projection $P_\Sigma : \mathbb{R}^N \to \Sigma$ that has the restricted Lipschitz property, we have $\beta_\Sigma(P_\Sigma) > 1$.*

Proof By contradiction, assume that $P_\Sigma : \mathbb{R}^N \to \Sigma$ has the 1-Lipschitz property. Let $z \in \mathbb{R}^N \setminus \Sigma$. From Lemma 3.3, $Q_1(P_\Sigma(z), z, x)$ is upper bounded with respect to x only if for all $x \in \Sigma$, $\langle P_\Sigma(z) - z, x\rangle = 0$. However, since Span$(\Sigma) = \mathbb{R}^N$, this extends to $\forall w \in \mathbb{R}^N$: just write $w = \sum_i \lambda_i x_i$ with $x_i \in \Sigma$. Then

$$\langle P_\Sigma(z) - z, w\rangle = \langle P_\Sigma(z) - z, \sum_i \lambda_i x_i\rangle = \sum_i \lambda_i \langle P_\Sigma(z) - z, x_i\rangle = 0. \tag{56}$$

Take $w = P_\Sigma(z) - z$, we deduce that $\|P_\Sigma(z) - z\|_2^2 = 0$ and $z = P_\Sigma(z) \in \Sigma$. Since we supposed $z \notin \Sigma$, we reach a contradiction. □

We show in the following Lemma that while orthogonal projections onto unions of subspaces are not linear in general, they are homogeneous.

Lemma 3.6 (Homogeneity of orthogonal projections on unions of subspaces) *Let $\Sigma \subset \mathbb{R}^N$ be a (potentially infinite) union of subspaces. Let $\lambda \in \mathbb{R}$, $z \in \mathbb{R}^N$. Let $(x_t)_{t\geq 0}$ a sequence of elements $x_t \in \Sigma$ such that $\|x_t - z\|_2 \to_{t\to\infty} \inf_{x\in\Sigma} \|x - z\|_2$, then $\|\lambda x_t - \lambda z\|_2 \to_{t\to\infty} \inf_{x\in\Sigma} \|x - \lambda z\|$.*

In particular, if the orthogonal projection onto Σ exists, we have

$$P_\Sigma^\perp(\lambda z) = \lambda P_\Sigma^\perp(z). \tag{57}$$

Proof For $z = 0$, remark that $P_\Sigma^\perp(0) = 0$ always exists for union of subspaces.

For $z \neq 0$. Let $(x_t)_{t\geq 0}$ as defined in the hypotheses. We have that, for all $\varepsilon > 0$, there is t_0 such that $t > t_0$ implies

$$\|x_t - z\|_2 \leq \inf_{x\in\Sigma} \|x - z\|_2 + \varepsilon. \tag{58}$$

Let $\lambda \in \mathbb{R}$, we have

$$\begin{aligned} \|\lambda x_t - \lambda z\|_2 = |\lambda| \|x_t - z\|_2 &\leq |\lambda| \inf_{x\in\Sigma} \|x - z\|_2 + |\lambda|\varepsilon \\ &= \inf_{x\in\Sigma} |\lambda| \|x - z\|_2 + |\lambda|\varepsilon \\ &= \inf_{x\in\Sigma} \|\lambda x - \lambda z\|_2 + |\lambda|\varepsilon \end{aligned} \tag{59}$$

As Σ is homogeneous, $\inf_{x\in\Sigma} \|\lambda x - \lambda z\|_2 = \inf_{\tilde{x}\in\Sigma} \|\tilde{x} - \lambda z\|_2$ and

$$\|\lambda x_t - \lambda z\|_2 = |\lambda| \|x_t - z\|_2 \leq \inf_{x\in\Sigma} \|\tilde{x} - \lambda z\|_2 + |\lambda|\varepsilon. \tag{60}$$

As we can find such t_0, for any ε, we just showed that

$$\|\lambda x_t - \lambda z\|_2 \xrightarrow[t\to\infty]{} \inf_{x\in\Sigma} \|\tilde{x} - \lambda z\|_2. \tag{61}$$

If $x = P_\Sigma^\perp(z)$ exists, the infimum is indeed a minimum reached at some $x^\star \in \Sigma$ and $\lambda x^\star = \arg\min_{x\in\Sigma} \|x - \lambda z\|_2$. □

As we saw, the Lipschitz property of the projection P_Σ is *necessary* for the general proof of convergence of the algorithm with a linear rate. In the following, we show that there is always a projection (the orthogonal projection provided its existence) having this property for general unions of subspaces.

Theorem 3.1 *Let Σ be a (potentially infinite) union of subspaces such that the orthogonal projection onto Σ exists then $P_\Sigma^\perp$ has restricted Lipschitz constant $\beta = \sqrt{\frac{5+\sqrt{13}}{2}} \approx 2.07$.*

Proof Let $c = \beta^2 \geq 2$. Let $z \in \mathbb{R}^N$ and $W \subset \Sigma$ a subspace such that $P_W^\perp(z) \in P_\Sigma^\perp(z)$.

Let $V \subset \Sigma$ such that $P_V^\perp(P_W^\perp(z) - cz) \in P_\Sigma^\perp(P_W^\perp(z) - cz)$.

Using the expression of Q_c in (49), with the homogeneity of the orthogonal projection 3.6, we have

$$\begin{aligned} Q_c(P_\Sigma^\perp(z), z, x^\star) &= \|P_\Sigma^\perp(z)\|_2^2 - c\|z\|_2^2 + \frac{1}{c-1}\|P_\Sigma^\perp(P_\Sigma^\perp(z) - cz)\|_2^2 \\ &= \|P_W^\perp(z)\|_2^2 - c\|z\|_2^2 + \frac{1}{c-1}\|P_V^\perp(P_W^\perp(z) - cz)\|_2^2 \\ &= \|P_W^\perp(z)\|_2^2 - c\|z\|_2^2 + \frac{1}{c-1}\|P_V^\perp(P_W^\perp(z) - z - (c-1)z)\|_2^2. \end{aligned} \tag{62}$$

We develop

$$\begin{aligned}
&Q_c(P_\Sigma^\perp(z),z,x^\star)\\
&=\|P_W^\perp(z)\|_2^2-c\|z\|_2^2\\
&\quad+\frac{1}{c-1}\left(\|P_V^\perp(P_W^\perp(z)-z)\|_2^2+(c-1)^2\|P_V^\perp(z)\|_2^2-2(c-1)\langle P_V^\perp(P_W^\perp(z)-z),P_V^\perp(z)\rangle\right)\\
&=(1-c)\|P_W^\perp(z)\|_2^2-c\|P_W^\perp(z)-z\|_2^2\\
&\quad+\frac{1}{c-1}\|P_V^\perp(P_W^\perp(z)-z)\|_2^2+(c-1)\|P_V^\perp(z)\|_2^2-2\langle P_V^\perp(P_W^\perp(z)-z),P_V^\perp(z)\rangle.
\end{aligned}\tag{63}$$

We have, as $I-P_W^\perp=P_{W^\perp}^\perp$,

$$\begin{aligned}
-\langle P_V^\perp(P_W^\perp(z)-z),P_V^\perp(z)\rangle&=\langle z-P_W^\perp(z),P_V^\perp(z)\rangle\\
&=\langle z-P_W^\perp(z),P_V^\perp(z)-z\rangle+\langle z-P_W^\perp(z),z\rangle\\
&=\langle z-P_W^\perp(z),P_V^\perp(z)-z\rangle+\langle P_{W^\perp}^\perp z,z\rangle\\
&=\langle z-P_W^\perp(z),P_V^\perp(z)-z\rangle+\langle P_{W^\perp}^\perp z,P_{W^\perp}^\perp z\rangle\\
&=\langle z-P_W^\perp(z),P_V(z)-z\rangle+\|z-P_W^\perp(z)\|_2^2.
\end{aligned}\tag{64}$$

With the Cauchy-Schwarz inequality,

$$\begin{aligned}
-\langle P_V^\perp(P_W^\perp(z)-z),P_V^\perp(z)\rangle&\leq\|z-P_W^\perp(z)\|_2\|P_V^\perp(z)-z\|_2+\|z-P_W^\perp(z)\|_2^2\\
&=\|z-P_W^\perp(z)\|\sqrt{\|z\|_2^2-\|P_V^\perp(z)\|_2^2}+\|z-P_W^\perp(z)\|_2^2.
\end{aligned}\tag{65}$$

We deduce from (63):

$$\begin{aligned}
&Q_c(P_\Sigma^\perp(z),z,x^\star)\\
&\leq(1-c)\|P_W^\perp(z)\|_2^2+(2-c)\|P_W^\perp(z)-z\|_2^2+\frac{1}{c-1}\|P_V^\perp(P_W^\perp(z)-z)\|_2^2\\
&\quad+(c-1)\|P_V^\perp(z)\|_2^2+2\|z-P_W^\perp(z)\|\sqrt{\|z\|_2^2-\|P_V^\perp(z)\|_2^2}.
\end{aligned}\tag{66}$$

Let us define the function $F(u)=2\|z-P_W^\perp(z)\|_2\sqrt{\|z\|_2^2-u}+(c-1)u$. We have

$$\begin{aligned}
&Q_c(P_\Sigma^\perp(z),z,x^\star)\\
&\leq(1-c)\|P_W^\perp(z)\|_2^2+(2-c)\|P_W^\perp(z)-z\|_2^2+\frac{1}{c-1}\|P_V^\perp(P_W^\perp(z)-z)\|_2^2+F\left(\|P_V^\perp(z)\|_2^2\right).
\end{aligned}\tag{67}$$

We have $F'(u)=-\frac{\|z-P_W^\perp(z)\|_2}{\sqrt{\|z\|_2^2-u}}+(c-1)$ and $F'(u^\star)=0$ if $\|z\|_2^2-u^\star=\frac{\|z-P_W^\perp(z)\|_2^2}{(c-1)^2}$ i.e. $u^\star=\|z\|_2^2-\frac{\|z-P_W^\perp(z)\|_2^2}{(c-1)^2}\geq 0$ (for $c\geq 2$). We have $F'(0)=(c-1)-\frac{\|z-P_W^\perp(z)\|_2}{\|z\|_2}\geq 0$ (as $\frac{\|z-P_W^\perp(z)\|_2}{\|z\|_2}\leq 1$ and $c\geq 2$)

and $F'(\|z\|_2^2) = -\infty$, we deduce that F is increasing from 0 to $u^\star$ and decreasing from $u^\star$ to $\|z\|_2^2$. We maximize $F(u)$ for $0 \leq u = \|P_V^\perp(z)\|_2^2 \leq \|P_W^\perp(z)\|_2^2$ (where the last inequality is guaranteed by the properties of the orthogonal projection, see Lemma 3.2).

We deduce that,

- if $u^\star \geq \|P_W^\perp(z)\|_2^2$, we have

$$\max_{0\leq u\leq \|P_W^\perp(z)\|_2^2} F(u) = F(\|P_W^\perp(z)\|_2^2) = 2\|z - P_W^\perp(z)\|_2^2 + (c-1)\|P_W(z)\|_2^2; \tag{68}$$

- if $0 \leq u^\star \leq \|P_W^\perp(z)\|_2^2$,

$$\max_{0\leq u\leq \|P_W^\perp(z)\|_2^2} F(u) = F(u^\star) = 2\frac{\|z - P_W^\perp(z)\|_2^2}{c-1} + (c-1)u^\star \tag{69}$$

As $u^\star \leq \|P_W^\perp(z)\|_2^2$ and $c \geq 2$,

$$\max_{0\leq u\leq \|P_W^\perp(z)\|_2^2} F(u) \leq 2\|z - P_W(z)\|_2^2 + (c-1)\|P_W^\perp(z)\|_2^2. \tag{70}$$

We deduce that overall, $F\left(\|P_V^\perp(z)\|_2^2\right) \leq 2\|z - P_W^\perp(z)\|_2^2 + (c-1)\|P_W^\perp(z)\|_2^2$ and

$$\begin{aligned} Q_c(P_\Sigma^\perp(z), z, x^\star) &\leq (1-c)\|P_W^\perp(z)\|_2^2 - c\|P_W^\perp(z) - z\|_2^2 \\ &\quad + \frac{1}{c-1}\|P_V^\perp(P_W^\perp(z) - z)\|_2^2 + 2\|P_W^\perp(z) - z\|_2^2 + (c-1)\|P_W^\perp(z)\|_2^2 \\ &\quad + 2\|P_W^\perp(z) - z\|_2^2) \\ &= -c\|P_W^\perp(z) - z\|_2^2 + \frac{1}{c-1}\|P_V^\perp(P_W^\perp(z) - z)\|_2^2 + 4\|P_W^\perp(z) - z\|_2^2 \end{aligned} \tag{71}$$

Using the fact that $\|P_V^\perp(P_W^\perp(z) - z)\|_2^2 \leq \|P_W^\perp(z) - z\|_2^2$, we have

$$\begin{aligned} Q_c(P_\Sigma^\perp(z), z, x^\star) &\leq -c\|P_W^\perp(z) - z\|_2^2 + \frac{1}{c-1}\|P_W^\perp(z) - z\|_2^2 + 4\|P_W^\perp(z) - z\|_2^2 \\ &= \left(\frac{4(c-1) + 1 - c(c-1)}{c-1}\right)\|P_W^\perp(z) - z\|_2^2 \\ &= \frac{-c^2 + (1+4)c + 1 - 4}{c-1}\|P_W^\perp(z) - z\|_2^2 \\ &= \frac{-c^2 + 5c - 3}{c-1}\|P_W^\perp(z) - z\|_2^2 \end{aligned} \tag{72}$$

The sign of Q_c is equal to the sign of $-c^2 + 5c - 3$. The determinant of this quadratic polynomial in c is $\Delta = 5^2 - 12 = 13 > 0$ leading to the largest square root $\frac{5+\sqrt{\Delta}}{2} = \frac{5+\sqrt{13}}{2}$. We deduce that $Q_c \leq 0$ for

$$c \geq \frac{5+\sqrt{13}}{2}. \tag{73}$$

This implies that $Q_\beta^2(P_\Sigma^\perp(z), z, x^\star) \leq 0$ for $\beta = \sqrt{\frac{5+\sqrt{13}}{2}}$ and $P_\Sigma^\perp$ is restricted β-Lipschitz. □

This result shows the global linear convergence of projected gradient descent with orthogonal projection for a large class of low-dimensional models. For more particular model sets such as sparse models, we can give a better estimation of the restricted Lipschitz constant which leads to an optimality result (when considering a collection of models) of the orthogonal projection.

3.2. *An optimality result for sparse recovery with iterative hard thresholding*

We observe that iterative hard thresholding for sparse recovery fits well with the previous framework. Indeed, for sparse recovery ($\Sigma = \Sigma_k$ the set of vectors with at most k non-zero elements), $P_\Sigma^\perp(z) = \mathrm{HT}(z)$ where HT is the hard thresholding operator selecting k largest absolute amplitudes in z.

The restricted Lipschitz property of the hard thresholding operator is a direct corollary of Theorem 3.1. We give a tighter restricted Lipschitz constant with a dedicated proof with the following theorem.

Theorem 3.2 (Restricted Lipschitz property of hard thresholding) *Let $\Sigma = \Sigma_k$. Then $P_\Sigma^\perp$ has the restricted Lipschitz condition w.r.t to Σ with constant $\beta = \sqrt{\frac{3+\sqrt{5}}{2}} \approx 1.618$*

Proof We use the characterization of the Lipschitz constant with the function Q_c given in Lemma 3.1. Let $c > 1$. Let $z \in \mathbb{R}^N$. We write z_S the restriction of z to a support S. Let $z_T = P_\Sigma^\perp(z) = \mathrm{HT}(z)$, with a support T that selects k largest amplitudes in z.

We use Lemma 3.3. The maximum of Q_c with respect to x is reached at

$$\begin{aligned} x^\star &= P_\Sigma^\perp\left(\frac{1}{1-c}(z_T - cz)\right) \\ &= P_\Sigma^\perp\left(z_T + \frac{c}{c-1} z_{T^c}\right). \end{aligned} \tag{74}$$

We define I the set of (less than k) coordinates of $z_T + \frac{c}{1-c} z_{T^c}$ selected by $P_\Sigma^\perp$ in T and J the coordinates selected in T^c. Note that $|I| + |J| = k$. We have

$$\begin{aligned} Q_c(P_\Sigma^\perp(z), z, x^\star) &= Q_c(z_T, z, x^\star) \\ &= (1-c)\|z_T\|_2^2 - c\|z_{T^c}\|_2^2 + (c-1)\|P_\Sigma^\perp\left(z_T + \frac{cz_{T^c}}{1-c}\right)\|_2^2 \\ &= (1-c)\|z_T\|_2^2 - c\|z_{T^c}\|_2^2 + (c-1)\|z_I\|_2^2 + \frac{c^2}{c-1}\|z_J\|_2^2 \\ &= (1-c)\|z_{T\setminus I}\|_2^2 - c\|z_{T^c}\|_2^2 + \frac{c^2}{c-1}\|z_J\|_2^2 \end{aligned} \tag{75}$$

We remark that $|T \setminus I| = k - |I| = |J|$. By definition of T any coordinate of $z_{T\setminus I}$ is larger than a coordinate of z_J as $J \subset T^c$. We deduce $\|z_J\|_2^2 \leq \|z_{T\setminus I}\|_2^2$ and, as $1-c>0$,

$$\begin{aligned} Q_c(P_\Sigma^\perp(z), z, x^\star) &\leq (1-c)\|z_J\|_2^2 - c\|z_J\|_2^2 + \frac{c^2}{c-1}\|z_J\|_2^2 \\ &= (1-2c+\frac{c^2}{c-1})\|z_J\|_2^2 \\ &= \frac{-c^2+3c-1}{c-1}\|z_J\|_2^2 \end{aligned} \tag{76}$$

The last expression is zeroed for $c = c^\star := 3/2 + \sqrt{5}/2 \approx 2.62$, (indeed $-(3/2+\sqrt{5}/2)^2 + 9/2 + 3\sqrt{5}/2 - 1 = 9/4 - 1 - 5/4 - 3\sqrt{5}/2 + 3\sqrt{5}/2 = 0$) i.e. for any $z \in \mathbb{R}^N, x \in \Sigma$

$$Q_{c^\star}(P_\Sigma^\perp(z), z, x) \leq 0. \tag{77}$$

and $P_\Sigma^\perp$ is restricted β - Lipschitz with $\beta = \sqrt{c^\star}$. $\square$

Going back to our condition we have $\delta^2\beta^2 < 1$ i.e. linear recovery with rate $r = \delta^2\beta^2$ provided A has a RIC on $\Sigma - \Sigma = \Sigma_{2k}$ with constant

$$\delta < \frac{1}{\beta} = \frac{1}{\sqrt{c^\star}} \approx 0.618 \tag{78}$$

(the threshold for recovery with convex methods and potentially sublinear rates is $\delta < \frac{1}{\sqrt{2}} \approx 0.707$ [36]). In [13], hard thresholding pursuit (with a linear rate of convergence) is successful if $\delta_{3k} < 0.57$. As the RIP on Σ_{3k} is much more stringent than the RIP on Σ_{2k}, we conclude that our general result is state-of-the-art for sparse recovery.

In the following, we show that the restricted Lipschitz constant given by Theorem 3.2 is tight when considering the collection of all sparse models. We also show that the orthogonal projection is a projection having the best Lipschitz constant if we consider the whole collection of sparse models. For a fixed sparse model the orthogonal projection is near-optimal.

Theorem 3.3 (An optimality result for hard thresholding) *For any $k \geq 1$, consider $\Sigma = \Sigma_k$ the k-sparse model set. Let Π_{Σ_k} the set projections onto Σ_k with a restricted Lipschitz property. Let*

$$\beta_k^\star = \inf_{P \in \Pi_{\Sigma_k}} \beta_{\Sigma_k}(P). \tag{79}$$

Then,

1. *the restricted Lipschitz constant from Theorem 3.2 is tight when considering the collection of sparse models for all sparsities $k \geq 1$, i.e.*

$$\sup_{k\geq 1} \beta_{\Sigma_k}^2(P_{\Sigma_k}^\perp) = \frac{3+\sqrt{5}}{2}; \tag{80}$$

2. *for any* $k \geq 3$

$$2.457 \approx \frac{3+\sqrt{11/3}}{2} \leq (\beta_k^\star)^2 \leq \beta^2_{\Sigma_k}(P^\perp_{\Sigma_k}) \leq \frac{3+\sqrt{5}}{2} \approx 2.618. \tag{81}$$

3. *we have the optimality of the orthogonal projection with respect to the sequence of models* Σ_k*:*

$$\sup_{k\geq 3}(\beta_k^\star)^2 = \sup_{k\geq 3}\beta^2_{\Sigma_k}(P^\perp_{\Sigma_k}) = \frac{3+\sqrt{5}}{2}. \tag{82}$$

Proof Recall that the maximisation of Q_c from Lemma 3.1 with respect to $x \in \Sigma$ yields

$$R_c(u,z) = Q_c(u,z,x^\star) = \|u\|_2^2 - c\|z\|_2^2 + \frac{1}{c-1}\|P^\perp_\Sigma(u-cz)\|_2^2. \tag{83}$$

The idea of this proof is to optimize R_c for a specific choice of z, yielding necessary conditions on c (a lower bound) to obtain $R_c(u^\star,z) = Q_c(u^\star,z,x^\star(z)) \leq 0$.

Let $z \in \mathbb{R}^N$ defined by $z_i = 1$ for $1 \leq i \leq k+1$ and $z_i = 0$ for $i > k+1$. We will show that $u^\star = P^\perp_\Sigma(z) = \mathrm{HT}(z)$ is a minimizer of $R_c(u,z)$ with respect to $u \in \Sigma$. We first show that $R_c(\cdot,z)$ is minimized by $u \in \Sigma$ such that $\mathtt{supp}(u) \subset \{1,\dots,k+1\}$.

Restriction of the support of $u^\star$**.** Let $u \in \Sigma$. Let $I = |\mathtt{supp}(u)| \cap \{1,\dots,k+1\}$ and $J = |\mathtt{supp}(u)| \setminus I$. We have

$$\begin{aligned} R_c(u_I,z) - R_c(u,z) &= \|u_I\|_2^2 + \frac{1}{c-1}\|P^\perp_\Sigma(u_I - cz)\|_2^2 \\ &\quad - \|u_I+u_J\|_2^2 - \frac{1}{c-1}\|P^\perp_\Sigma(u_I+u_J-cz)\|_2^2. \\ &= -\|u_J\|_2^2 + \frac{1}{c-1}\|P^\perp_\Sigma(u_I-cz)\|_2^2 - \frac{1}{c-1}\|P^\perp_\Sigma(u_I+u_J-cz)\|_2^2. \end{aligned} \tag{84}$$

As $J \cap \mathtt{supp}(u_I - cz) = \emptyset$, we have that $\|P^\perp_\Sigma(u_I-cz)\|_2^2 = \|\mathrm{HT}(u_I-cz)\|_2^2 \leq \|\mathrm{HT}(u_I+u_J-cz)\|_2^2 = \|P^\perp_\Sigma(u_I+u_J-cz)\|_2^2$. We deduce that

$$R_c(u_I,z) - R_c(u,z) \leq -\|u_J\|_2^2 \leq 0 \tag{85}$$

We deduce that we can consider $u^\star$ minimizing $R_c(u,z)$ such that $|\mathtt{supp}(u^\star)| \subset \{1,\dots,k+1\}$, i.e

$$\min_{u\in\Sigma} R_c(u,z) = \min_{u\in\Sigma,\mathtt{supp}(u)\subset\mathtt{supp}(z)} R_c(u,z). \tag{86}$$

Explicit minimization of $R_c(\cdot,z)$**.** Let $u \in \Sigma$ such that $\mathtt{supp}(u) \subset \{1,\dots,k+1\}$, e.g. (without loss of generality) $\mathtt{supp}(u) \subset \{1,\dots,k\}$ as z is constant on $\{1,\dots,k+1\}$.

We distinguish two cases:

- **Case 1:** for all $i \in \{1,\dots,k\}, |u_i - cz_i| = |u_i - c| \geq c$. Then, $P^\perp_\Sigma(u-cz) = \mathrm{HT}(u-cz) = (u_1 - cz_1,\dots,u_k-cz_k,0\dots0)$ and we have

$$R_c(u,z) = \sum_{i=1}^k |u_i|^2 - c(k+1) + \frac{1}{c-1}\sum_{i=1}^k |u_i - c|^2 \geq k\frac{c^2}{c-1} - (k+1)c. \tag{87}$$

- **Case 2:** there exists $j \in \{1,\dots,k\}$ such that $|u_j - c| < c$. Without loss of generality (just reorder the supports), suppose $|u_k - c| = \min_{i\in\{1,\dots,k\}} |u_i - c|$. In this case, we have that $P_\Sigma^\perp(u - cz) = \mathrm{HT}(u - cz)$ necessarily selects index $k+1$ (because $cz_{k+1} = c > |u_j - c|$). We deduce that $\|P_\Sigma^\perp(u-cz)\|_2^2 = c^2 + \sum_{i=1}^{k-1} |u_i - c|^2$.

  Then

$$R_c(u,z) = \sum_{i=1}^{k} |u_i|^2 - c(k+1) + \frac{1}{c-1}(c^2 + \sum_{i=1}^{k-1} |u_i - c|^2). \tag{88}$$

  We minimize $R_c(u,z)$ with respect to u_i, $i \neq k$ and the constraint $|u_i - c| \geq |u_k - c| =: \lambda$. We have

$$\arg \min_{u_i, |u_i - c| \geq \lambda} R_c(u,z) = \arg \min_{u_i, |u_i - c| \geq \lambda} |u_i|^2 + \frac{1}{c-1}|u_i - c|^2 =: g(u_i). \tag{89}$$

  remarking that $g'(u_i) = 2u_i + \frac{2}{c-1}(u_i - c) = 0$ for $u_i = 1$. This second-degree polynomial is minimized at $u_i^\star = 1$ if $|c-1| \geq \lambda = |u_k - c|$ (the global minimum is within the constraint). In this case, this gives

$$R_c((u_1^\star, \dots, u_{k-1}^\star, u_k), z) = (k-1) + |u_k|^2 + \frac{1}{c-1}(c^2 + (k-1)(1-c)^2) - c(k+1) \tag{90}$$

  Minimizing $|u_k|^2$ with respect to u_k such that $|u_k - c| \leq |u_i^\star - c| = |c-1|$ yields $u_k^\star = 1$ (The function is minimized on the constraint as the global minimum 0 is not on the constraint). We deduce that

$$\begin{aligned} R_c(u^\star, z) &= k + \frac{1}{c-1}(c^2 + (k-1)(c-1)^2) - c(k+1) \\ &= (c-1)(k-1) + k - c(k+1) + \frac{c^2}{c-1} \\ &= (c-1)k - (c-1) + (1-c)k - c + \frac{c^2}{c-1} = \frac{c^2}{c-1} - 2c + 1; \end{aligned} \tag{91}$$

  i.e.

$$\min_{u:\mathrm{supp}(u)\subset\{1,\dots,k\} \min_{i\in\{1,\dots,k\}} |u_i - c| \leq c-1} R_c(u,z) = \frac{c^2}{c-1} - 2c + 1. \tag{92}$$

  Now, we consider the case where u is such that $|u_k - c| > |c-1|$. Starting from (88), we minimize R_c with respect to u_i, $i \neq k$. This yields $u_i^\star = u_k$ (as the function g from (89) is minimized on the constraint) and

$$R_c((u_1^\star, \dots, u_{k-1}^\star, u_k), z) = k|u_k|^2 + \frac{1}{c-1}(c^2 + (k-1)|u_k - c|^2) - c(k+1) =: F(u_k). \tag{93}$$

  We have $F'(u_k) = 2ku_k + 2\frac{k-1}{c-1}(u_k - c)$ and $F'(u_k^\star) = 0$ if $(k(c-1) + (k-1))u_k^\star = (k-1)c$, i.e. if $(kc-1)u_k^\star = (k-1)c$ and $u_k^\star = \frac{(k-1)c}{kc-1} = \frac{(k-1)c}{(k-1)c+c-1} \leq 1$. We deduce that the inequality $|c - u_k^\star| \leq c$ is verified and

$$R_c(u^\star, z) = F\left(\frac{(k-1)c}{(k-1)c+c-1}\right), \tag{94}$$

i.e.

$$\min_{u:\mathrm{supp}(u)\subset\{1,\dots,k\}, c>\min_{i\in\{1,\dots,k\}}|u_i-c|\geq c-1} R_c(u,z) = F\left(\frac{(k-1)c}{(k-1)c+c-1}\right). \tag{95}$$

Now, with (91), remark that

$$\min_{u:\mathrm{supp}(u)\subset\{1,\dots,k\}, \min_{i\in\{1,\dots,k\}}|u_i-c|\leq c-1} R_c(u,z) = F(1) = \frac{c^2}{c-1} - 2c + 1. \tag{96}$$

With (95) and (96), we get

$$\begin{aligned}\min_{u:\mathrm{supp}(u)\subset\{1,\dots,k\}, \min_{i\in\{1,\dots,k\}}|u_i-c|<c} R_c(u,z) &= \min\left(F(1), F\left(\frac{(k-1)c}{(k-1)c+c-1}\right)\right)\\ &= F\left(\frac{(k-1)c}{(k-1)c+c-1}\right).\end{aligned} \tag{97}$$

We compare with the lower bound from (87), we have, for $k > 2$

$$\begin{aligned}k\frac{c^2}{c-1} - (k+1)c - F(1) &= (k-1)\frac{c^2}{c-1} - (k-1)c - 1\\ &= (k-1)c\left(\frac{c}{c-1} - 1\right) - 1\\ &= (k-1)\frac{c}{c-1} - 1 \geq k-2 > 0.\end{aligned} \tag{98}$$

We deduce that for $k > 2$

$$\begin{aligned}\min_{u:\mathrm{supp}(u)\subset\{1,\dots,k\}} R_c(u,z) &= \min_{u:\mathrm{supp}(u)\subset\{1,\dots,k\}, \min_{i\in\{1,\dots,k\}}|u_i-c|<c} R_c(u,z)\\ &= F\left(\frac{(k-1)c}{(k-1)c+c-1}\right).\end{aligned} \tag{99}$$

Best uniform constant independent of k. We have shown that

$$\sup_{\tilde{z}\in\mathbb{R}^N} \min_{u\in\Sigma} R_c(u,\tilde{z}) \geq \min_{u\in\Sigma} R_c(u,z) = F\left(\frac{(k-1)c}{(k-1)c+c-1}\right) \to_{k\to\infty} F(1). \tag{100}$$

Moreover as $F(1) = \frac{-c^2+3c-1}{c-1}$, for $c \geq 1$, $F(1) \leq 0$ if and only if $c \geq \frac{3+\sqrt{5}}{2} \approx 2.6180$.

We have that $P_\Sigma^\star \in \arg\min_{P\in\Pi_\Sigma} \beta_\Sigma(P)$ implies

$$\sup_{\tilde{z}\in\mathbb{R}^N} R_c(P_\Sigma^\star(\tilde{z}),\tilde{z}) = \sup_{\tilde{z}\in\mathbb{R}^N} \min_{u\in\Sigma} R_c(u,\tilde{z}) \geq F\left(\frac{(k-1)c}{(k-1)c+c-1}\right). \tag{101}$$

Using Theorem 3.2, we have that, for any s

$$\begin{aligned}\frac{3+\sqrt{5}}{2} &\geq \sup_{k\geq 1} \beta_{\Sigma_k}(P_{\Sigma_k}^\perp)^2 \\ &\geq \sup_{k\geq 3} \beta_{\Sigma_k}(P_{\Sigma_k}^\star)^2 \sup_{k\geq 3} \min_{c\geq 1; F\left(\frac{(k-1)c}{(k-1)c+c-1}\right)\leq 0} c = \min_{c\geq 1; F(1)\leq 0} c = \frac{3+\sqrt{5}}{2}.\end{aligned} \tag{102}$$

This proves the third conclusion of this Theorem. Note that this also forces the equality $\sup_{k\geq 1} \beta_{\Sigma_k}(P_{\Sigma_k}^\perp)^2 = \frac{3+\sqrt{5}}{2}$, which is the first conclusion of this theorem.

Lower bound with $k=3$. For $k=3$, we have, using the definition of F from (93),

$$\begin{aligned}F\left(\frac{(k-1)c}{(k-1)c+c-1}\right) &= F\left(\frac{2c}{3c-1}\right) \\ &= 3\left(\frac{2c}{3c-1}\right)^2 + \frac{1}{c-1}(c^2+2|\frac{2c}{3c-1}-c|^2) - 4c \\ &= c\left(3c\left(\frac{2}{3c-1}\right)^2 + \frac{c}{c-1}(1+2|\frac{2}{3c-1}-1|^2) - 4\right) \\ &= c\left(3c\left(\frac{2}{3c-1}\right)^2 + \frac{c}{c-1}(1+2\left(\frac{-3c+3}{3c-1}\right)^2) - 4\right)\end{aligned} \tag{103}$$

We reduce to the same denominator.

$$\begin{aligned}F\left(\frac{(k-1)c}{(k-1)c+c-1}\right) &= c\left(\frac{12c(c-1)+c(3c-1)^2+2c(-3c+3)^2-4(3c-1)^2(c-1)}{(3c-1)^2(c-1)}\right) \\ &= c\left(\frac{12c(c-1)+c(3c-1)^2+18c(c-1)^2-4(3c-1)^2(c-1)}{(3c-1)^2(c-1)}\right) \\ &= c\left(\frac{(c-1)(12c+18c(c-1))+c(3c-1)^2-4(3c-1)^2(c-1)}{(3c-1)^2(c-1)}\right) \\ &= c\left(\frac{(c-1)(18c^2-6c)+c(3c-1)^2-4(3c-1)^2(c-1)}{(3c-1)^2(c-1)}\right) \\ &= c\left(\frac{6c(c-1)(3c-1)+c(3c-1)^2-4(3c-1)^2(c-1)}{(3c-1)^2(c-1)}\right) \\ &= c\left(\frac{6c(c-1)+c(3c-1)-4(3c-1)(c-1)}{(3c-1)(c-1)}\right) = c\frac{-3c^2+9c-4}{(3c-1)(c-1)}.\end{aligned} \tag{104}$$

Calculating the roots of the second degree polynomial in c given by $-3c^2+9c-4$, we deduce that $F\left(\frac{(k-1)c}{(k-1)c+c-1}\right) \leq 0$ for $c \geq \frac{3+\sqrt{11/3}}{2} \approx 2.457$.

Hence, for $k \geq 3$,

$$2.457 \leq \beta^2_{\Sigma_k}(P^\star_{\Sigma_k}) \leq \beta^2_{\Sigma_k}(P^\perp_{\Sigma_k})^2 \approx 2.618. \tag{105}$$

This is the second conclusion of this Theorem. □

3.3. *Discussion*

We have shown that for generic models (union of subspaces), the orthogonal projection plays an important role within the set of possible projections onto Σ. In particular, it is nearly optimal for sparse recovery (we conjecture the same result for low-rank recovery, as it is straightforward to extend our proofs to this case). More surprisingly, for a fixed sparsity (or an arbitrary union of subspaces), the orthogonal projection might not be optimal. Our investigations reveal that it minimizes $Q_c(u,z,x^\star)$ only for some $z \in \mathbb{R}^N$. However the bound on Lipschitz constants shows that there is little to be gained with another projection in the case of sparse recovery.

Another important conclusion that we can draw from these results, is that for general unions of subspaces, we can always bound reasonably the restricted Lipschitz constant of the orthogonal projection. Finding optimal projections for a general model is still an open question. In particular, does the orthogonal projection have the same near-optimality result for an arbitrary union of subspaces as iterative hard thresholding for sparse recovery?

4. Application to inverse problems with deep priors

We have shown in the previous section that projected gradient descent identifies low-dimensional models with a linear rate as soon as a projection P_Σ with the restricted Lipschitz condition (Definition 2.1) on the low-dimensional model Σ is available. In this Section, we show that the plug-and-play (PnP) framework can be interpreted as a low-dimensional model recovery and that, experimentally, for image inverse problems, linear rates of convergence to the underlying low-dimensional model are observed. This shows that the restricted Lipschitz condition appears to hold approximately in practice and that global faster rates are obtained beyond the typical non-convex setting of the literature.

4.1. *An interpretation of the plug-and-play method with low-dimensional recovery theory*

Many variations of PnP methods exist in the literature [22, 41]. PnP methods use a general denoiser to approximate the proximal operator associated with a regularization function, which could be interpreted as an operator minimizing a distance between a low-dimensional model and a given point, i.e. a projection operator. In the following we use the formalism of the proximal gradient method [41] (PnP-PGM), which directly uses the denoiser in a projected gradient descent scheme and yields state of the art results for inverse problems in imaging.

Let $f(x) = \frac{1}{2}\|Ax-y\|_2^2$. In this context, the problem of estimating $\hat{x}$ from $y = A\hat{x}$ is solved using :

$$x_{n+1} = D(x_n - \mu A^H(Ax_n - y)) \tag{106}$$

where D is the general purpose denoiser and μ is the gradient step size. We fall in the iterations defined in Theorem 2.1 (with projection and descent steps reversed). Hence *global* convergence will be guaranteed if D is a projection onto a set Σ (which is exactly the set of fixed points of the denoiser D)

with *restricted* Lipschitz constant $\beta_\Sigma(D)$. Note that, to the best of our knowledge convergence of PnP methods rely on a much more stringent *global* Lipschitz condition on the considered objects (objective functional, etc, ..., see the Related work section 1.4). Hence, only sublinear rates are obtained. We show that experimentally the convergence to the underlying low-dimensional model set Σ (the fixed points of the denoiser D) corresponds to linear rates. We will illustrate this linear convergence in the following experimental sections.

Note that we conjectured in the previous sections that, as PnP-PGM can be interpreted as a projected gradient descent, this method should have better rates than the other natural choice of regularizing direction $(I-D)(x_n)$ (see Section 2.2) which was indeed proposed in the gradient method regularization by denoising (GM-RED) [31] :

$$x_{n+1} = x_n - \mu(A^H(Ax_n - y) + \lambda(I-D)(x_n)). \tag{107}$$

We also propose a comparison of the two methods in Section 4.2.3 in light of our results.

4.2. *Experiments*

In practice, PnP-PGM and GM-RED never fully recover $\hat{x}$ due to approximations error. Indeed, the unknown $\hat{x}$ is never a perfect fixed point of the denoiser D and thus we cannot expect it to be fully recovered. Further approximation errors may also occur which makes it complicated for the theory to perfectly fit. Hence, it is near impossible for the sequence $(x_n)_{n\geq 1}$ given by (106) to have a linear convergence rate with respect to $\hat{x}$ as stated by Theorem 2.1. However, under the assumption that x_n has a r-linear rate of convergence with respect to $x^* = \lim_{n\to+\infty} x_n$, we have

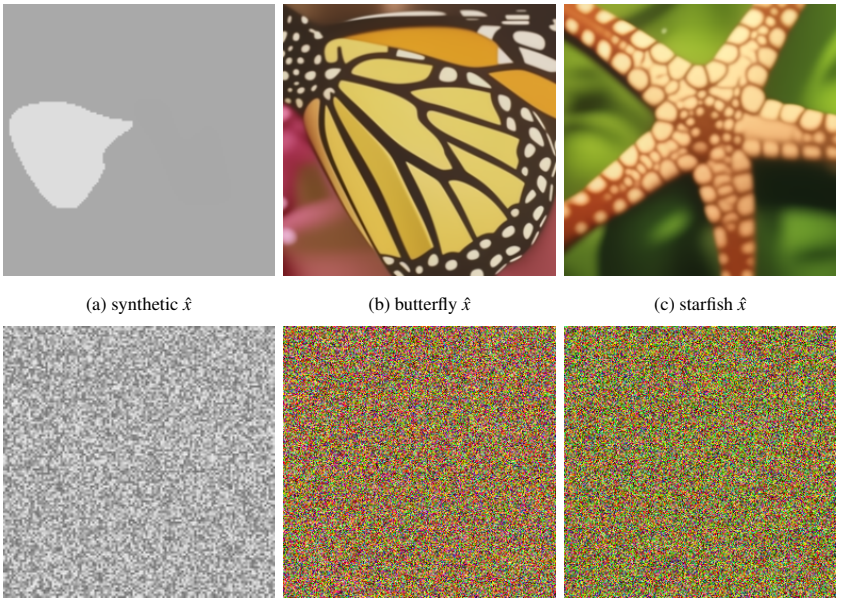
$$\begin{aligned} \|x_n - \hat{x}\|_2^2 &\leq (\|x_n - x^*\| + \|x^* - \hat{x}\|)^2 \\ &\leq (r^{\frac{n}{2}}\|x_0 - x^*\| + \|x^* - \hat{x}\|)^2. \end{aligned} \tag{108}$$

This illustrates the fact that in case of linear convergence to an approximate $\hat{x}$ without knowing the exact limit x^*, we should observe a convergence given by the last inequality in (108).

In the following experiments, we compare this estimation of convergence with other sublinear convergence rates $\frac{1}{n}$, $\frac{1}{n^2}$ on both synthetic images and natural images. We conduct our experiments on two different linear operators A: a mask operator that erases 30% of the pixels of the image and a Gaussian blur operator (with $\sigma = 1.0,\ 3.0$). For each algorithm, the tuning parameters μ and λ are selected through line search such that they ensure the best recovery of $\hat{x}$ (independently for each method). Moreover, the initialization x_0 is set using a random uniform distribution and the mean is reset such that it matches the target image mean. The code and weights of the denoisers used for the numerical experiments can be found in the open-access GitLab repository [35]. We show in Figure 1, the synthetic and natural images and initializations used in our experiments.

4.2.1. Synthetic piecewise-constant images

In this experiment, we use a denoiser D (without parameter for the noise level), parametrized with a DRUNet architecture [41] (a state-of-the-art combination of ResNet and U-Net architectures, without explicit low-dimensional latent space), trained on a dataset of randomly generated piecewise-constant images (see [17] for a precise description of the random synthetic dataset). The target image $\hat{x}$ (see Fig. 1a) was generated such that it had a high fixed point value with respect to the denoiser ($\mathrm{PSNR}(\hat{x}, D(\hat{x})) \approx 64.84$). We observe that the theoretical bound (108) matches well with the observed convergence rate



(a) synthetic $\hat{x}$ (b) butterfly $\hat{x}$ (c) starfish $\hat{x}$

(d) initialization x_0 (e) initialization x_0 (f) initialization x_0

FIG. 1. Target images used in numerical experiments. For each numerical experiment, the image $\hat{x}$ is set such that it is approximately a fixed point of the respective denoiser. (a) The synthetic image is generated such that it is an approximate of a fixed point of the denoiser. (b-c) The DRUNet denoiser is applied a few times on a natural image such that we obtain an approximation of a fixed point of the denoiser. (d-f) The initialization for each image respectively, generated from a random uniform distribution and with the same mean as the respective target image.

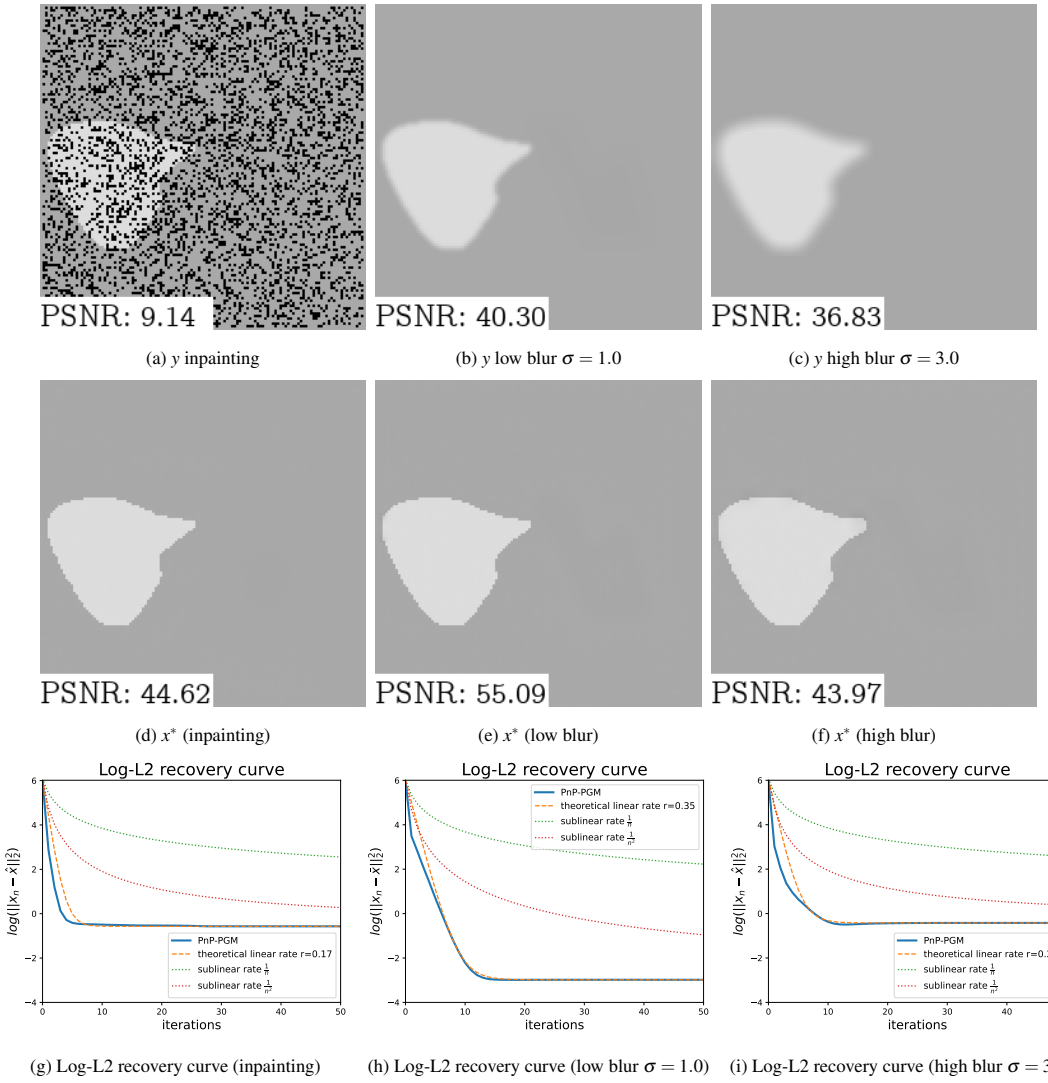
$\|x_n - \hat{x}\|$ in Figure 2, whereas the sublinear rates $\frac{1}{n}$ and $\frac{1}{n^2}$ did not. The theoretical rate in Figure 2 is calculated as the minimal rate upper-bounding the experimental convergence curve.

4.2.2. Natural Images

In these experiments, we use a DRUNet denoiser trained on natural images (we used the weights provided by the DeepInverse library, see Acknowledgements at the end of the paper) and with the noise level η as input (non-blind denoising). To obtain a fixed point of the denoiser, we apply the denoiser on an original natural image multiple times with the entry noise level set to η=0.18 (This parameter is manually set to give good performance). For both images tested (see Figures 3 and 4), we observe that the theoretical linear convergence rate curve (calculated in the same way as the previous experiment) indeed matches well with the convergence curve $\|x_n - \hat{x}\|_2^2$. As expected, the higher blur (i.e. a degraded Restricted Isometry Constant) leads to a decrease in the performance of recovery. Also note that when the linear rate of convergence decreases as in the high blur experiment of Figure 4, it becomes harder to distinguish from a fast sub-linear rate ($\frac{1}{n^2}$).

4.2.3. Comparison between PNP-PGM and GM-RED

We perform the same previous experiments with the GM-RED algorithm, which falls outside the theoretical results provided in this paper. Once again, we observe that the theoretical linear convergence rate matched very well the convergence curve $\|x_n - \hat{x}\|_2^2$ (see Fig. 5, 6, 7). Furthermore, the measured convergence rate of GM-RED is slower than PnP-PGM, reinforcing our initial conjecture. Interestingly,



(a) y inpainting (b) y low blur $\sigma = 1.0$ (c) y high blur $\sigma = 3.0$

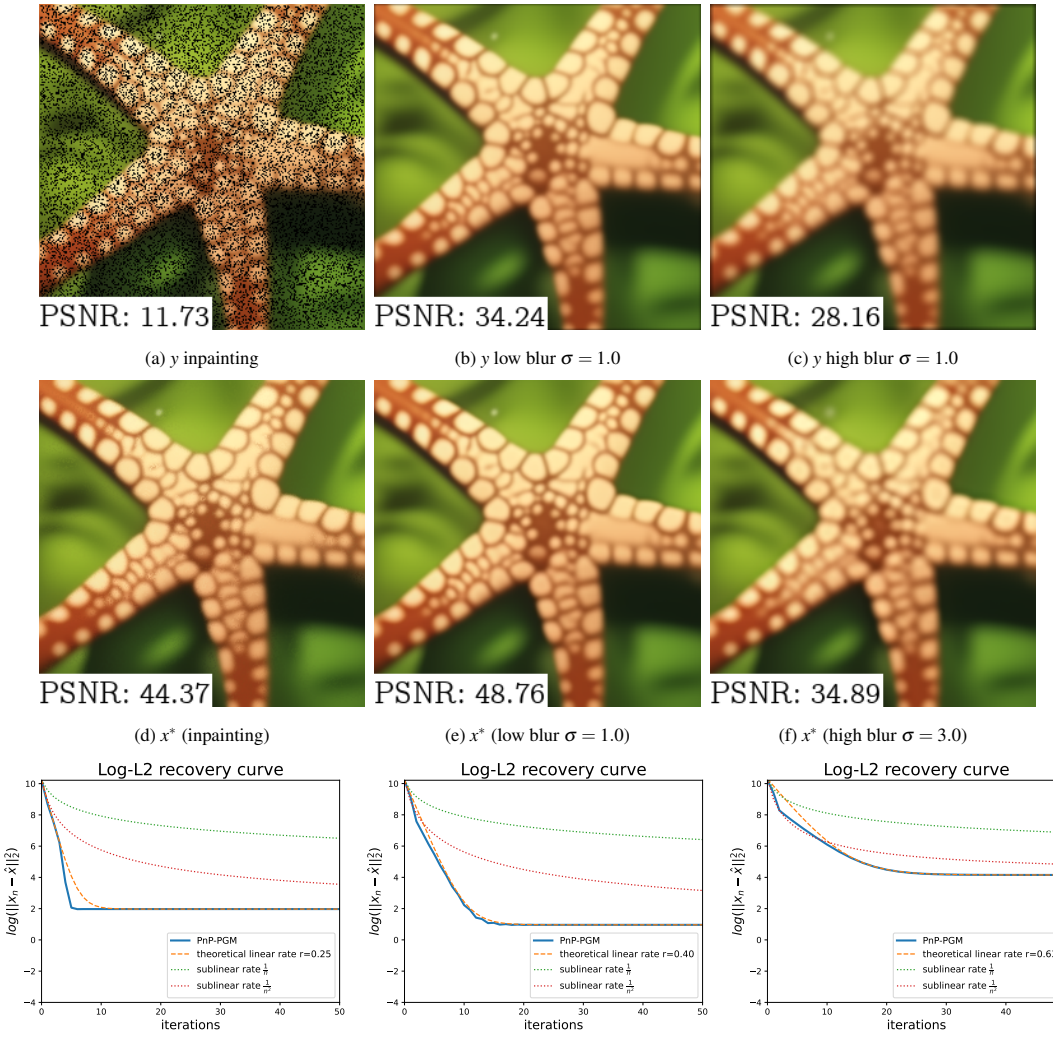
(d) x^* (inpainting) (e) x^* (low blur) (f) x^* (high blur)

(g) Log-L2 recovery curve (inpainting) (h) Log-L2 recovery curve (low blur $\sigma = 1.0$) (i) Log-L2 recovery curve (high blur $\sigma = 3.0$)

FIG. 2. Experiments for the PnP-PGM algorithm (106) on a synthetic image that is an approximate fixed point of a DRUNet denoiser. For each linear measurement operator, the theoretical linear convergence rate matches well the Log-L2 recovery curve.

we found that GM-RED was able to obtain a better recovery of our estimated fixed point than PnP-PGM in some experiments (particularly when the measurement operator is better conditioned (low blur)). This could be due to the additional parameter in GM-RED which allows a finer direction towards the low-dimensional model Σ induced by the denoiser or simply that our choice of approximate fixed point just matches better the GM-RED algorithm. This opens a question to include approximation error in the projection in our analysis (as was for example done in [15]).

While these experiments do not prove the optimality (with respect to the rate of convergence) of the projected gradient method for low-dimensional recovery, it suggests that our analysis lays out good foundations for a global understanding of low-dimensional recovery with deep priors.



(a) y inpainting (b) y low blur $\sigma = 1.0$ (c) y high blur $\sigma = 1.0$

(d) x^* (inpainting) (e) x^* (low blur $\sigma = 1.0$) (f) x^* (high blur $\sigma = 3.0$)

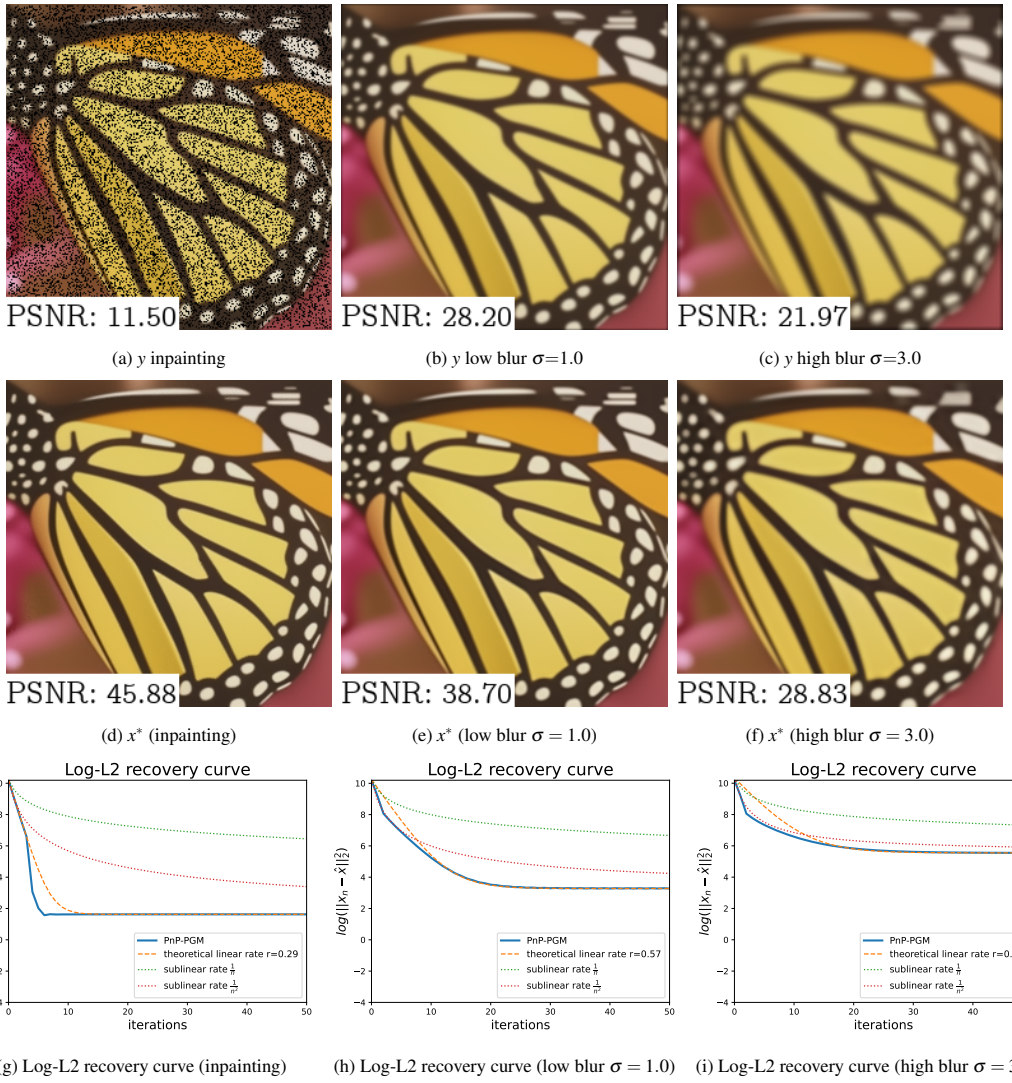
(g) Log-L2 recovery curve (inpainting) (h) Log-L2 recovery curve (low blur $\sigma = 1.0$) (i) Log-L2 recovery curve (high blur $\sigma = 3.0$)

FIG. 3. Experiment of the PnP-PGM algorithm (106) on an image that is an approximate fixed point of the blind denoiser. For each linear operator, the theoretical linear convergence rate matches well the Log-L2 recovery curve.

5. Conclusion

We have given a convergence analysis of a class of projected descent algorithms for the recovery of low-dimensional models. This class of algorithm appears naturally when proving the fast convergence of the wider class of algorithms of averaged directions. Our result explicitly quantifies the convergence rate with the restricted isometry constants of the measurement operator and a newly introduced restricted Lipschitz condition on the operator projecting onto the model set. This decouples the role of the geometry of the model and the quality of the measurement operator in the rate of convergence.

More particularly we have shown that the orthogonal projection yields very general guarantees for unions of subspaces. These guarantees can be improved in the case of sparse recovery and iterative



(a) y inpainting (b) y low blur σ=1.0 (c) y high blur σ=3.0

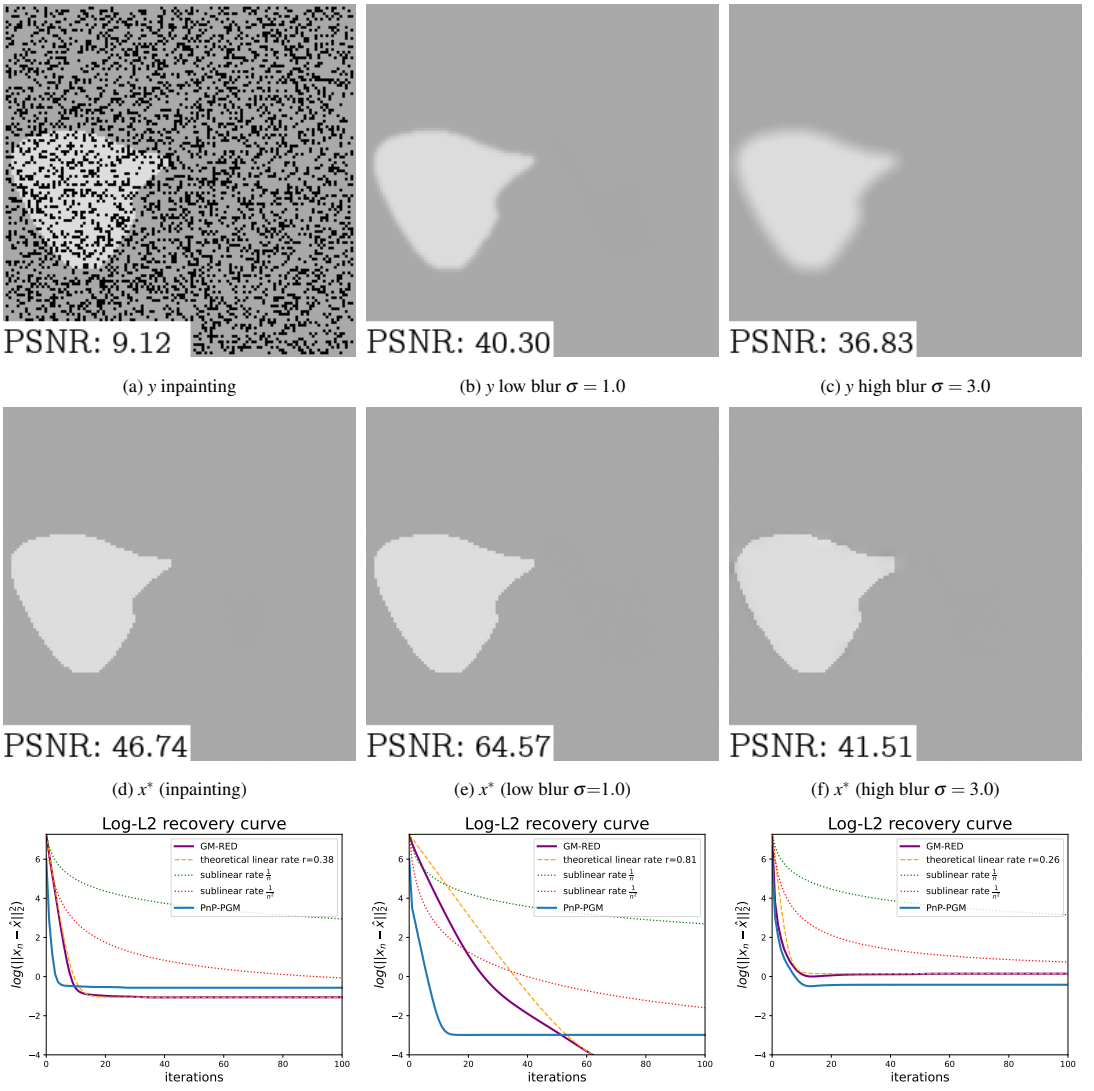
(d) x^* (inpainting) (e) x^* (low blur $\sigma = 1.0$) (f) x^* (high blur $\sigma = 3.0$)

(g) Log-L2 recovery curve (inpainting) (h) Log-L2 recovery curve (low blur $\sigma = 1.0$) (i) Log-L2 recovery curve (high blur $\sigma = 3.0$)

FIG. 4. Experiment of the PnP-PGM algorithm (106) on an image that is an approximate fixed point of the blind denoiser. For each linear operator, the theoretical linear convergence rate matches well the Log-L2 recovery curve.

hard thresholding, showing that hard thresholding is indeed optimal for the convergence rate (via the restricted Lipschitz constant) when considering the whole class of sparse models (for any sparsity).

Our work lays out the foundation of a theoretical framework for optimal algorithms for the recovery of low-dimensional beyond the variational approach. Many ideas can be explored to generalize this work. Extending to more general classes of algorithms, exploring the tightness of our different results, or studying the impact of the noise in the search for optimality are possible interesting leads. Another possibility would be to add more flexibility to the uniform rate condition and to consider a finite time "burning" period, i.e. linear convergence guaranteed after a fixed number of iterations.



(a) y inpainting (b) y low blur $\sigma = 1.0$ (c) y high blur $\sigma = 3.0$

(d) x^* (inpainting) (e) x^* (low blur σ=1.0) (f) x^* (high blur $\sigma = 3.0$)

(g) Log-L2 recovery curve (inpainting) (h) Log-L2 recovery curve (low blur $\sigma = 1.0$) (i) Log-L2 recovery curve (high blur $\sigma = 3.0$)

FIG. 5. Experiments of the GM-RED algorithm (106) on a synthetic image for different measurement operators. We observe that GM-RED presents a linear convergence rate. Moreover, the measured convergence rate is slower than PnP-PGM.

Our results guarantee linear convergence for solving inverse problems with deep priors. They also raise the question of learning a projection (a denoiser in the plug-and-play framework) with a good *restricted* Lipschitz constant, thus relaxing the global Lipschitz condition.

6. Acknowledgements

Experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil

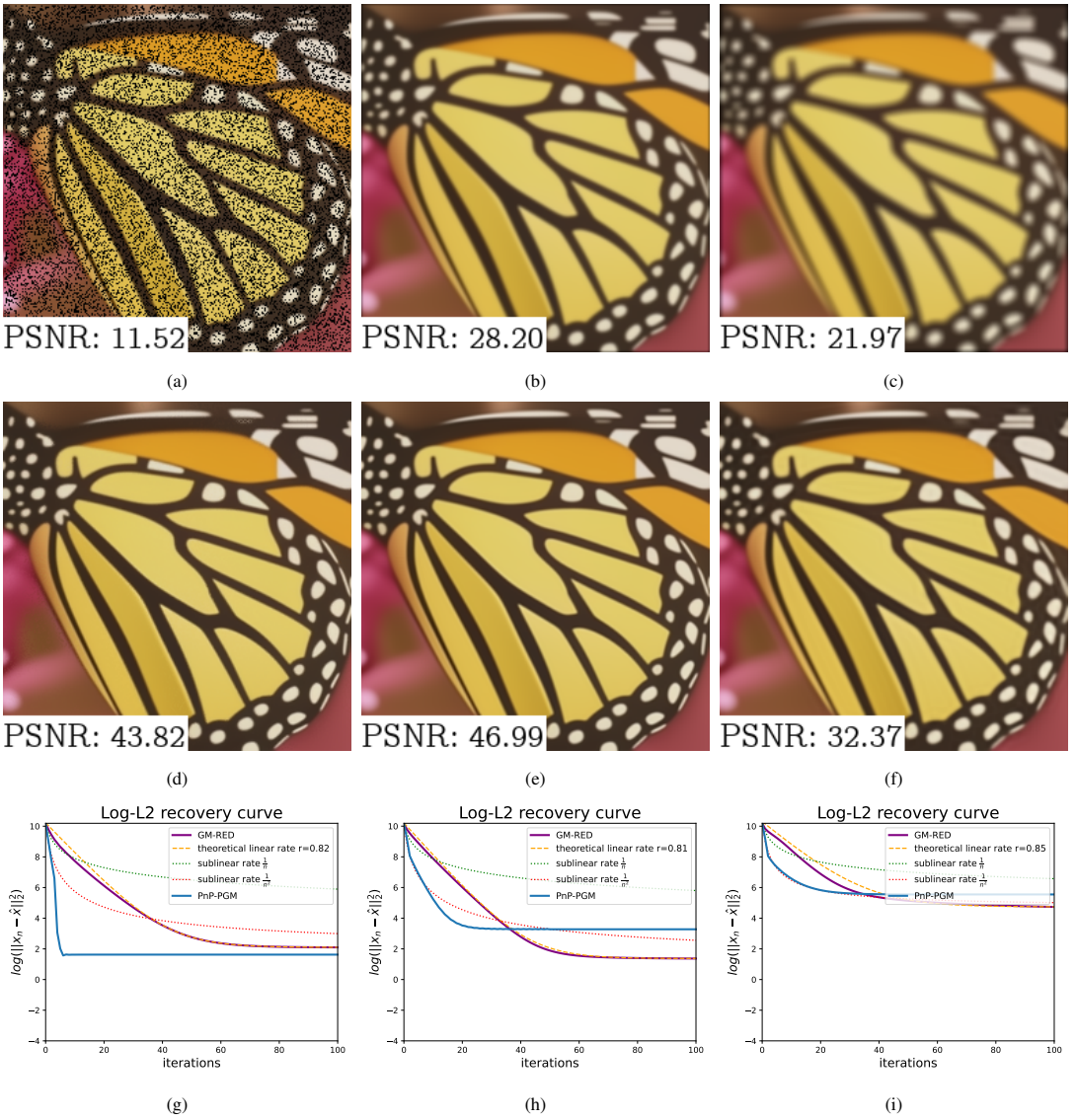
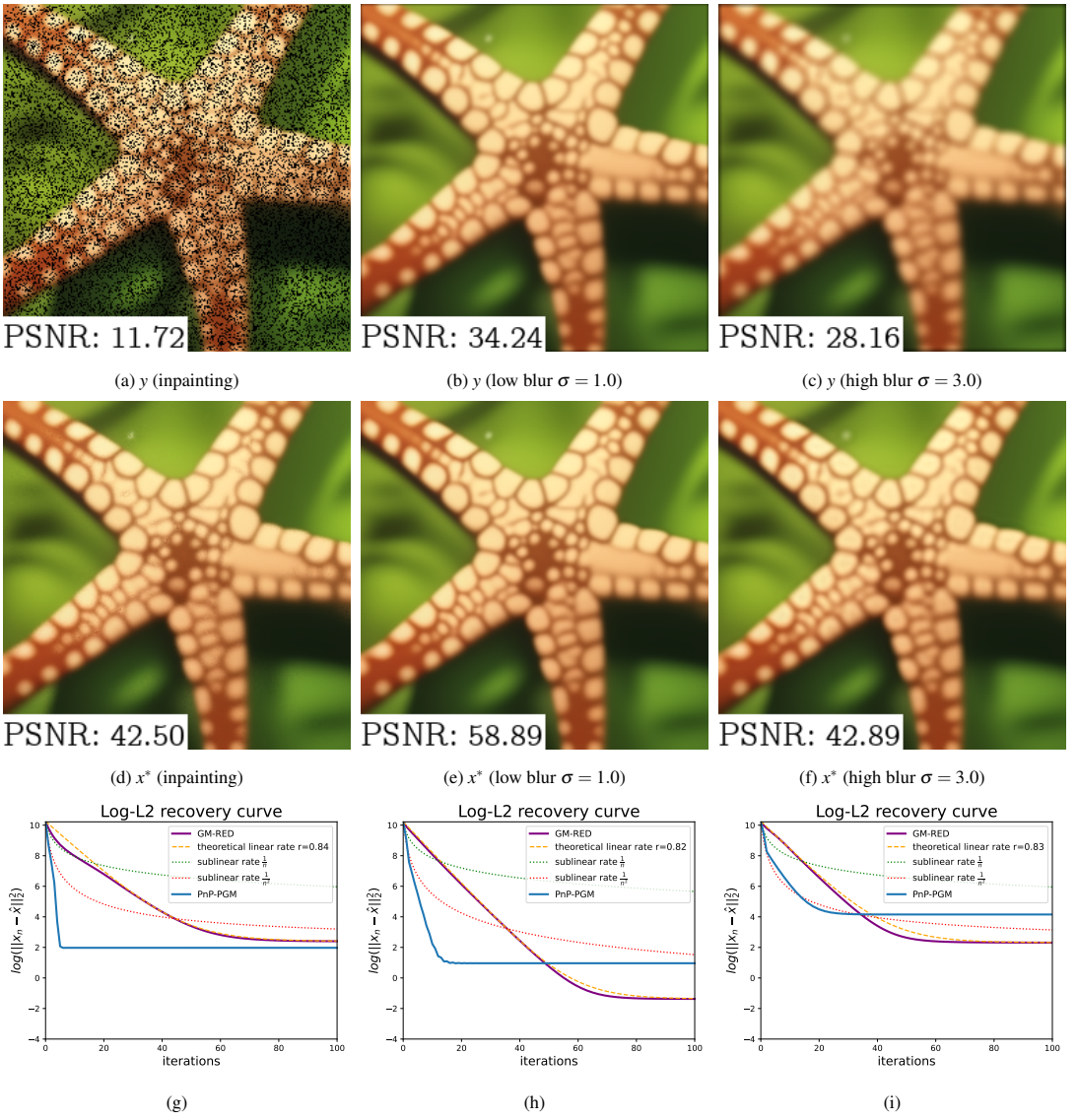


FIG. 6. Experiment of the GM-RED algorithm (106) on the butterfly image for different linear operations. We observe that GM-RED presents a linear convergence rate. Moreover, the measured convergence rate is slower than PnP-PGM.

Régional d'Aquitaine (see https://www.plafrim.fr). Furthermore, we are grateful to the DeepInverse python library (https://deepinv.github.io/deepinv/index.html) from which the code and weights of the denoiser for natural images was taken from. This work was supported by the French National Research Agency (ANR) under reference ANR-20-CE40-0001 (EFFIREG project), and by PEPR PDE_AI.



(a) y (inpainting) (b) y (low blur $\sigma = 1.0$) (c) y (high blur $\sigma = 3.0$)

(d) x^* (inpainting) (e) x^* (low blur $\sigma = 1.0$) (f) x^* (high blur $\sigma = 3.0$)

(g) (h) (i)

FIG. 7. Experiment of the GM-RED algorithm (106) on the starfish image for different linear operations. We observe that GM-RED presents a linear convergence rate. Moreover, the measured convergence rate is slower than PnP-PGM.

REFERENCES

1. 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*, Mathematical Programming, 137 (2013), pp. 91–129.
2. S. BAHMANI, P. T. BOUFOUNOS, AND B. RAJ, *Learning model-based sparsity via projected gradient descent*, IEEE Transactions on Information Theory, 62 (2016), pp. 2092–2099.
3. R. F. BARBER AND W. HA, *Gradient descent with non-convex constraints: local concavity determines*

convergence, Information and Inference: A Journal of the IMA, 7 (2018), pp. 755–806.
4. A. BASTOUNIS, A. C. HANSEN, AND V. VLAČIĆ, *The extended smale's 9th problem–on computational barriers and paradoxes in estimation, regularisation, computer-assisted proofs and learning*, arXiv preprint arXiv:2110.15734, (2021).
5. A. BECK AND N. HALLAK, *Optimization problems involving group sparsity terms*, Mathematical Programming, 178 (2019), pp. 39–67.
6. T. BLUMENSATH AND M. E. DAVIES, *Normalized iterative hard thresholding: Guaranteed stability and performance*, IEEE Journal of selected topics in signal processing, 4 (2010), pp. 298–309.
7. A. BOURRIER, M. E. DAVIES, T. PELEG, P. PÉREZ, AND R. GRIBONVAL, *Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems*, IEEE Transactions on Information Theory, 60 (2014), pp. 7928–7946.
8. P. BÜRGISSER AND F. CUCKER, *Condition: The geometry of numerical algorithms*, vol. 349, Springer Science & Business Media, 2013.
9. W. CHEN, D. WIPF, AND M. RODRIGUES, *Deep learning for linear inverse problems using the plug-and-play priors framework*, in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 8098–8102.
10. R. COHEN, M. ELAD, AND P. MILANFAR, *Regularization by denoising via fixed-point projection (red-pro)*, SIAM Journal on Imaging Sciences, 14 (2021), pp. 1374–1406.
11. F. CUCKER AND S. SMALE, *Complexity estimates depending on condition and round-off error*, Journal of the ACM (JACM), 46 (1999), pp. 113–184.
12. M. A. DAVENPORT AND J. ROMBERG, *An overview of low-rank matrix recovery from incomplete observations*, IEEE Journal of Selected Topics in Signal Processing, 10 (2016), pp. 608–622.
13. S. FOUCART, *Hard thresholding pursuit: an algorithm for compressive sensing*, SIAM Journal on numerical analysis, 49 (2011), pp. 2543–2563.
14. S. FOUCART AND H. RAUHUT, *A Mathematical Introduction to Compressive Sensing*, Springer, 2013.
15. M. GOLBABAEE AND M. E. DAVIES, *Inexact gradient projection and fast data driven compressed sensing*, IEEE Transactions on Information Theory, 64 (2018), pp. 6707–6721.
16. M. GONZÁLEZ, A. ALMANSA, AND P. TAN, *Solving inverse problems by joint posterior maximization with autoencoding prior*, SIAM Journal on Imaging Sciences, 15 (2022), pp. 822–859.
17. A. GUENNEC, J.-F. AUJOL, AND Y. TRAONMILIN, *Joint structure-texture low dimensional modeling for image decomposition with a plug and play framework*, (2024).
18. P. HAND AND V. VORONINSKI, *Global guarantees for enforcing deep generative priors by empirical risk*, IEEE Transactions on Information Theory, 66 (2019), pp. 401–418.
19. A. HAUPTMANN, S. MUKHERJEE, C.-B. SCHÖNLIEB, AND F. SHERRY, *Convergent regularization in inverse problems and linear plug-and-play denoisers*, Foundations of Computational Mathematics, (2024), pp. 1–34.
20. X. HUO AND J. CHEN, *Complexity of penalized likelihood estimation*, Journal of Statistical Computation and Simulation, 80 (2010), pp. 747–759.
21. S. HURAULT, A. LECLAIRE, AND N. PAPADAKIS, *Gradient step denoiser for convergent plug-and-play*, in International Conference on Learning Representations (ICLR'22), 2022.
22. S. HURAULT, A. LECLAIRE, AND N. PAPADAKIS, *Gradient step denoiser for convergent plug-and-play*, (2022).
23. U. S. KAMILOV, C. A. BOUMAN, G. T. BUZZARD, AND B. WOHLBERG, *Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications*, IEEE Signal Processing Magazine, 40 (2023), pp. 85–97.
24. U. S. KAMILOV AND H. MANSOUR, *Learning optimal nonlinearities for iterative thresholding algorithms*, IEEE Signal Processing Letters, 23 (2016), pp. 747–751.
25. O. LEONG, E. O'REILLY, Y. S. SOH, AND V. CHANDRASEKARAN, *Optimal regularization for a data source*, arXiv preprint arXiv:2212.13597, (2022).
26. C. LI AND B. ADCOCK, *Compressed sensing with local structure: uniform recovery guarantees for the*

sparsity in levels class, Applied and Computational Harmonic Analysis, 46 (2019), pp. 453–477.
27. H. LIU AND R. FOYGEL BARBER, *Between hard and soft thresholding: optimal iterative thresholding algorithms*, Information and Inference: A Journal of the IMA, 9 (2020), pp. 899–933.
28. G. OLIKIER AND I. WALDSPURGER, *Projected gradient descent accumulates at bouligand stationary points*, arXiv preprint arXiv:2403.02530, (2024).
29. G. ONGIE, A. JALAL, C. A. METZLER, R. G. BARANIUK, A. G. DIMAKIS, AND R. WILLETT, *Deep learning techniques for inverse problems in imaging*, IEEE Journal on Selected Areas in Information Theory, 1 (2020).
30. P. PENG, S. JALALI, AND X. YUAN, *Solving inverse problems via auto-encoders*, IEEE Journal on Selected Areas in Information Theory, 1 (2020), pp. 312–323.
31. Y. ROMANO, M. ELAD, AND P. MILANFAR, *The little engine that could: Regularization by denoising (red)*, SIAM Journal on Imaging Sciences, 10 (2017), pp. 1804–1844.
32. V. ROULET, N. BOUMAL, AND A. D'ASPREMONT, *Computational complexity versus statistical performance on sparse recovery problems*, Information and Inference: A Journal of the IMA, 9 (2020), pp. 1–32.
33. J. SCARLETT, R. HECKEL, M. R. RODRIGUES, P. HAND, AND Y. C. ELDAR, *Theoretical perspectives on deep learning methods in inverse problems*, IEEE journal on selected areas in information theory, 3 (2022), pp. 433–453.
34. J. TACHELLA, D. CHEN, AND M. DAVIES, *Sensing theorems for unsupervised learning in linear inverse problems*, Journal of Machine Learning Research, 24 (2023), pp. 1–45.
35. Y. TRAONMILIN, J.-F. AUJOL, AND A. GUENNEC, *Linear rate numerical experiments*. https://plmlab.math.cnrs.fr/aguennec/beyond_variation_numerical_final, 2024.
36. Y. TRAONMILIN AND R. GRIBONVAL, *Stable recovery of low-dimensional cones in hilbert spaces: One rip to rule them all*, Applied and Computational Harmonic Analysis, 45 (2018), pp. 170–205.
37. Y. TRAONMILIN, R. GRIBONVAL, AND S. VAITER, *A theory of optimal convex regularization for low-dimensional recovery*, Information and Inference: A Journal of the IMA, 13 (2024), p. iaae013.
38. M. UNSER AND S. DUCOTTERD, *Parseval convolution operators and neural networks*, arXiv preprint arXiv:2408.09981, (2024).
39. S. V. VENKATAKRISHNAN, C. A. BOUMAN, AND B. WOHLBERG, *Plug-and-play priors for model based reconstruction*, in 2013 IEEE global conference on signal and information processing, IEEE, 2013, pp. 945–948.
40. T. VU AND R. RAICH, *On asymptotic linear convergence of projected gradient descent for constrained least squares*, IEEE Transactions on Signal Processing, 70 (2022), pp. 4061–4076.
41. K. ZHANG, Y. LI, W. ZUO, L. ZHANG, L. VAN GOOL, AND R. TIMOFTE, *Plug-and-play image restoration with deep denoiser prior*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44 (2021), pp. 6360–6376.