From sparse recovery to plug-and-play priors, understanding trade-offs for stable recovery with generalized projected gradient descent

Ali Joundi*, Yann Traonmilin*, Jean-François Aujol*

We consider the problem of recovering an unknown low-dimensional vector from noisy, underdetermined observations. We focus on the Generalized Projected Gradient Descent (GPGD) framework, which unifies traditional sparse recovery methods and modern approaches using learned deep projective priors. We extend previous convergence results to robustness to model and projection errors. We use these theoretical results to explore ways to better control stability and robustness constants. To reduce recovery errors due to measurement noise, we consider generalized back-projection strategies to adapt GPGD to structured noise, such as sparse outliers. To improve the stability of GPGD, we propose a normalized idempotent regularization for the learning of deep projective priors. We provide numerical experiments in the context of sparse recovery and image inverse problems, highlighting the trade-offs between identifiability and stability that can be achieved with such methods.

1. Introduction

We study the problem of recovering an unknown vector $\hat{x} \in \mathbb{R}^N$ from an undetermined number of noisy observations $y \in \mathbb{R}^m$ defined by:

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

where A is the measurement operator and $e \in \mathbb{R}^m$ is the measurement noise. Many problems in data science can be modeled this way, in particular, imaging problems, where $\hat{x}$ is an image (in biology, medicine, astronomy, ...).

As the number of measurements is often insufficient, $m < N$, a prior model on the unknown is *necessary*. In this article, we assume $\hat{x}$ lies approximately in Σ, where Σ is a *low-dimensional* model, i.e., a subset of $\mathbb{R}^N$ that can be described with few parameters. This setup is the basis of sparse recovery theory, where, under some conditions on the measurement operator A (e.g., number of measurements of random Gaussian matrices), it is possible to guarantee stable (with respect to noise) and robust recovery (with respect to model error) of elements of Σ with convex or non-convex algorithms. For example, for noise of finite energy, given an estimate $x^\star$ of $\hat{x}$, such guarantees are expressed,

$$\|x^\star - \hat{x}\|_2^2 \leq C_1 \|e\|_2^2 + C_2 d(\hat{x}, \Sigma) \tag{2}$$

where C_1, C_2 are constants and $d(\hat{x}, \Sigma)$ is some notion of distance to Σ.

Learning based methods, where the prior on the unknown vector $\hat{x}$ is learned, have been very successful at solving such inverse problems when a large database $X \subset \mathbb{R}^N$ of examples is available. A large part of the literature uses the minimization of potentially non-convex functionals to solve such problems. In particular, the popular plug-and-play methods use a general-purpose denoiser learned on X as a projective prior used for the minimization of such functionals and consider convergence to critical points of such functions. Many learning based methods can be understood as methods using *deep projective priors* where a generalized projection onto a non-explicit set Σ is learned. In [1], instead of considering the minimization of an underlying functional, it was proposed to

*Univ. Bordeaux, Bordeaux INP, CNRS, IMB, F-33400, Talence, France {ali.joundi,yann.traonmilin,jean-francois.aujol}@math.u-bordeaux.fr.

Third Conference on Parsimony and Learning (CPAL 2026).

consider unified guarantees of generalized projected gradient descent (GPGD). It was shown that GPGD for low-dimensional recovery can model, at the same time sparse recovery (through the classical iterative hard thresholding algorithm) and a class of learning based plug-and-play methods where the denoiser is used as a projection. We define GPGD iteration as

$$\mathcal{I}(P_\Sigma): \qquad x_{n+1} = P_\Sigma(x_n) - \mu A^T(AP_\Sigma(x_n) - y) \tag{3}$$

where P_Σ is a generalized projection on the low dimensional model Σ and $\mu > 0$ is a fixed step size. Note that projected gradient descent is often presented with the projection and descent step reversed. The "gradient" in (3) is the gradient of the ℓ^2-datafit $\nabla \frac{1}{2}\|Ax - y\|_2^2 = A^T(Ax - y)$. This gradient can also be interpreted as a back-projection of the residual $Ax - y$ to the ambient space of $\hat{x}$ through the back-projection operator A^T.

For sparse recovery, using the orthogonal projection leads to the iterative hard thresholding algorithm. In the context of plug-and-play methods, using a general-purpose denoiser as a projection leads to the proximal gradient descent plug-and-play method. Other methods relying on deep projective priors, such as auto-encoders, can be interpreted in this framework [2]. [1] shows that global linear recovery is possible provided the measurement operator A verifies a restricted isometry property and the projection verifies a *restricted Lipschitz property*. It is further shown that the restricted Lipschitz constant drives both the identifiability capabilities of Σ and the convergence rate. It is advocated that this constant is thus a good property to compare different GPGD algorithms to recover the same model and even to consider optimal methods (in this context, iterative hard thresholding is shown to be near-optimal for sparse recovery).

However, [1] only considers a perfect modeling without noise, model error and with "ideal" projection having the restricted Lipschitz property. [2] showed that stable recovery (stability to observation noise) was guaranteed under the same conditions. However, stability to noise has not been fully explored. For example, the considered PGD algorithm uses the gradient $A^T(Ax - y)$ of the ℓ^2 data-fit functional $\frac{1}{2}\|Ax - y\|_2^2$. While this gradient is well adapted to Gaussian white noise (as it is related to the maximum likelihood estimator), it might not be adapted to other types of degradations such as sparse corruptions, where the energy of the noise cannot be bounded. The validity of initial results with model error and approximate projections is still an open question.

In this paper, we study how the generalized projected gradient descent behaves under the presence of noise, modeling error, and approximate projection, in a framework that unifies sparse recovery and methods relying on deep projective priors. In particular, we study whether the restricted isometry and restricted Lipschitz conditions are sufficient to guarantee stable and robust recovery. We also discuss how to minimize or control different error terms in the design of GPGD methods.

1.1. Contributions

In Section 3, we provide a general stable and robust recovery theorem for generalized projected gradient descent with arbitrary back-projections. This theorem shows how using GPGD with general back-projections and approximate projections leads to stability to measurement noise, modeling error, and approximate projections. In such a context, the restricted isometry constants and restricted Lipschitz conditions are still the main factors impacting recovery.

In Section 4, when sparse outliers contaminate measurements, we illustrate how adapting the back-projection leads to stable recovery and provide experiments within the plug-and-play framework.

In Section 5, we propose a normalized idempotent regularization (NIPR) to control the approximate projection error. We show experimentally that NIPR for deep projective priors (plug-and-play priors and auto-encoder priors in Section F in the appendix) improves the stability of convergence of GPGD while preserving reconstruction quality.

1.2. Related work

There exists a wide body of work studying the projected gradient descent in various contexts. For sparse recovery, PGD is known as iterative hard thresholding, and its linear convergence under a

restricted isometry property of the operator A has been shown in [3, 4]. Similar results for function minimization with PGD without an explicit low-dimensional model were given in [5]. Approximate *orthogonal* projections for the recovery of low-dimensional models are studied in [6]. In [7], global convergence of PGD is given for a class of generalized sparsity models and the orthogonal projection. [1] decouples the convergence rate through a restricted isometry constant and a restricted Lpischitz constant of the projection. In this work, we will consider approximate *restricted Lipschitz* projections and stable recovery for generic noises. Global convergence of gradient projection has been shown under a general KL property in [8]. General stationary properties of the iterates of PGD are given in [9]. In this work, we only consider cases where linear convergence is proven.

Beyond the optimality approach in [1], other works intend to formally improve PGD algorithms. For thresholding algorithms, [10] optimizes a local concavity property to improve local convergence properties. In [11], an optimal non-linearity is learned from the data. In [12], a robust iterative hard thresholding algorithm is presented: it proposes to use $A^T W$ where W is a shrinkage operator calculated iteratively. However, no theoretical analysis is provided. In [13], an iterative thresholding applied solely to the residual is presented (no sparsity model on the data). In this work, we will discuss adaptation to noise in the context of outliers and the impact on recovery guarantees. Note that adaptation to noise in the variational context (minimization of a regularized data-fit functional) is generally done by adapting the norm of the data-fit (see e.g. [14–16] for sparse outliers).

The generalized projected gradient descent context allows us to study recovery algorithms relying on deep priors. Specifically, in the class of state-of-the-art plug-and-play (imaging) algorithms [17–20], the so-called proximal gradient method [21] is a generalized projected gradient descent algorithm where the projection is performed using a general-purpose denoiser learned with a deep neural network. While [1, 2] showed experimentally that this method linearly converges to approximate fixed points of the denoiser, a complete theoretical study of stability, robustness, and approximation of a projection with deep neural networks was not given.

In [22], in the context of generative modeling, the importance of having access to a real projection onto the data manifold is emphasized. In consequence, an idempotent regularization of the generative model is proposed (i.e $\|f \circ f - f\|$ where f is the generating function) and shows improved stability of the generation. We will build on this idea to improve the stability of deep projective priors. Note that there exists a recent wide literature on stochastic algorithms using generative models such as diffusion or flow matching priors to solve inverse problems (see e.g. surveys [23, 24]). While links with deep projective priors have been made in deterministic settings, such algorithms are out of the scope of this article where we focus on deterministic "plug-and-play like" algorithms.

2. Notations, definitions and previous results

We introduce our theoretical framework and recall previous results useful to understand our main theoretical result in the next section. We suppose that Σ is a homogeneous space (verified by sparse and low-rank models).

To guarantee uniform recovery, we assume that the measurement operator A verifies a restricted isometry property (a lower RIP is anyway *necessary* for the identifiability of Σ [25]). We use the following notion of restricted isometry constant.

Definition 2.1. *The operator B has a restricted isometry constant (RIC) $\delta < 1$ on the secant set $\Sigma - \Sigma = \{x_1 - x_2 : x_1, x_2 \in \Sigma\}$ if for all $x_1, x_2 \in \Sigma$*

$$\|(B - I)(x_1 - x_2)\|_2 \leq \delta \|x_1 - x_2\|_2 \tag{4}$$

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

Definition 2.1 is well adapted to the study of the composition of the measurement operator with a generalized back-projection of the form LA (where L is a linear operator) through the constant $\delta(LA)$ (in iterations (3), $L = A^T$). We consider the following notion of generalized projection.

Definition 2.2 (Generalized projection)**.** *Let $\Sigma \subset \mathbb{R}^N$. A (set-valued) generalized 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 $P(z)$. Considering such generalized projection brings flexibility to the framework: both sparse models and deep projective priors can be considered (as the latter might not be projections in the usual mathematical definition as idempotent operators $P \circ P = P$).

Definition 2.3 (Orthogonal projection)**.** *We define, when it exists, the (set-valued) orthogonal projection onto a set $\Sigma \subset \mathbb{R}^N$ as follows: for all $z \in \mathbb{R}^N$*

$$P_\Sigma^\perp(z) = \arg\min_{x \in \Sigma} \|x - z\|_2^2. \tag{5}$$

We recall the restricted Lipschitz property of the projection introduced by [1] to study linear convergence of GPGD.

Definition 2.4 (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 $z \in \mathbb{R}^N, x \in \Sigma, u \in P(z)$ we have*

$$\|u - x\|_2 \leq \beta \|z - x\|_2 \tag{6}$$

We denote by $\beta_\Sigma(P)$ the smallest β for which P satisfies the restricted β-Lipschitz property.

Note that for single-valued projection, this can be rewritten $\|P(z) - P(x)\|_2 \leq \beta \|z - x\|_2$, hence the name *restricted Lipschitz*. Linear recovery of $\hat{x}$ with GPGD has been shown when $\delta(A^T A)\beta_\Sigma(P) < 1$. If the orthogonal projection $P_\Sigma^\perp$ exists (Σ is then called *proximinal*), we have $1 \leq \beta_\Sigma(P_\Sigma^\perp) \leq 2$. It was shown that the optimal projection $P^\star$ minimizing $\beta_\Sigma(P^\star)$ exists. For sparse recovery ($\Sigma = \Sigma_k$, the set of sparse vectors), using the orthogonal projection corresponds to the iterative hard thresholding algorithm and is nearly optimal for the restricted Lipschitz constant with $\beta_{\Sigma_k}(P_{\Sigma_k}^\perp) \leq \sqrt{\frac{3+\sqrt{5}}{2}} \approx 1.618$. Also note that if $\beta_\Sigma(P) < \infty$ then P is necessarily idempotent.

3. Stable and robust linear recovery with GPGD with approximate projections

We give a general recovery result with GPGD with generalized back-projections and approximate projections. We include three deviations from an ideal noiseless model: stable recovery with generalized back projection, robustness to model error and robustness to approximate projections.

Adaptation to noise As was proposed in previous work [1], the restricted Lipschitz constant is a good objective to minimize in the search for optimal projections for GPGD algorithms given a low-dimensional model Σ. However, it is not clear in a noisy context what would be an optimal GPGD as we have guarantees of the form:

$$\|x_n - \hat{x}\|_2^2 \leq C_1 r^n + C_2 \|e\|_2 \tag{7}$$

where x_n are the iterates of the considered GPGD method and $C_1, C_2, r \geq 0$ are some constants.

To define an optimal GPGD method as minimizing such an upper bound, we need to optimize both the rate r and the stability constant C_2. When there is some knowledge about the structure of the noise, variational methods provide a way to adapt the recovery method by adapting the norm of the data fit term. In particular, it is known that in some cases, stability to unbounded noise (sparse corruptions) can be obtained using a sparsity-inducing norm such as the ℓ^1-norm for the data-fit term [14–16]. Following similar ideas, we propose to generalize the projected gradient descent by considering general back-projections instead of the back-projection A^T induced by a ℓ^2 data-fit.

Model error In practice, we have that $\hat{x} \notin \Sigma$, but we assume some control $d_{P_\Sigma}(\hat{x}, \Sigma) \leq \tau$ (or equivalently $\|\hat{x} - P_\Sigma(\hat{x})\|_2 \leq \tau$) for the distance to the model $d_{P_\Sigma}(\cdot, \Sigma)$ associated to the projection P_Σ.

Approximate projections In the plug-and-play framework, in GPGD, we use a general-purpose denoiser as P_Σ. While ensuring that P_Σ is actually a projection (i.e., it has a set of fixed points)

would guarantee linear convergence, in practice, it is not a real projection. In fact, it is often observed that iterations of PGD in this context diverge after having reached an optimal point. To model this effect, suppose that in place of a restricted β-Lipschitz projection P_Σ, we use a projection P such that:

$$P(x_n) = P_\Sigma(x_n) + R(x_n) \tag{8}$$

where $\|R(x_n)\|_2 \leq \eta$ for some constant $\eta > 0$.

We consider GPGD with generalized back-projection iterations:

$$\mathcal{I}_{GBP}(P, L): \qquad x_{n+1} = P(x_n) - \mu L(AP(x_n) - y) \tag{9}$$

where $L \in \mathbb{R}^{N \times m}$ is a general linear back-projection from the observation space to the ambient space of $\hat{x}$ and P_Σ is a generalized projection onto Σ. This way, the complexity of the back-projection step is limited to the cost of a matrix vector multiplication. We present our main convergence result.

Theorem 3.1. *Let $\Sigma \subset \mathbb{R}^N$. Let $\mu, \eta > 0$. Let P_Σ be a generalized projection onto Σ. Consider iterates from the GPGD with approximate projection $P = P_\Sigma + R$ and $\|R(x)\|_2 \leq \eta$ for all $x \in \mathbb{R}^N$. Suppose that μLA has restricted isometry constant $\delta := \delta(\mu LA)$ and P_Σ has restricted Lipschitz constant $\beta := \beta_\Sigma(P_\Sigma)$. Consider $d_{P_\Sigma}(\hat{x}, \Sigma) := \|P_\Sigma(\hat{x}) - \hat{x}\|_2$. Assuming $\delta\beta < 1$, we have*

$$\|x_n - P_\Sigma(\hat{x})\|_2 \leq (\delta\beta)^n \|x_0 - P_\Sigma(\hat{x})\|_2 + C_{\text{stab}}\|Le\|_2 + C_{\text{rob}} d_{P_\Sigma}(\hat{x}, \Sigma) + C_{\text{proj}}\eta \tag{10}$$

where we define the stability constant $C_{\text{stab}} := \frac{\mu}{1-\delta\beta}$; the robustness (to model error) constant $C_{\text{rob}} := \frac{1}{1-\delta\beta}\|\mu LA\|_{\text{op}}$; the approximate projection constant $C_{\text{proj}} := \frac{1}{1-\delta\beta}\|I - \mu LA\|_{\text{op}}$. We also have

$$\|x_n - \hat{x}\|_2 \leq (\delta\beta)^n \|x_0 - P_\Sigma(\hat{x})\|_2 + C_{\text{stab}}\|\mu Le\|_2 + C'_{\text{rob}} d_{P_\Sigma}(\hat{x}, \Sigma) + C_{\text{proj}}\eta \tag{11}$$

where $C'_{\text{rob}} := 1 + \frac{\|\mu LA\|_{\text{op}}}{1-\delta\beta}$.

This theorem shows linear convergence to a set of estimates having estimation error controlled by the noise level, the model error and the approximation of the projection with the constants C_{stab}, C_{rob} and C_{proj}, respectively. Particularly, in the specific case when $L = A^T$, $d_{P_\Sigma}(\hat{x}, \Sigma) = 0$, $\eta = 0$, we obtain recovery guarantees that were given in [1, 2].

With Theorem 3.1, we remark that for a given measurement operator A, the restricted Lipschitz constant of the projection P_Σ drives both the rate of convergence and identifiability of GPGD with backprojection $L \propto A^T$. The stability constant C_{stab}, the robustness constant C_{rob} and the approximate projection constant C_{proj} are also increasing with respect to $\beta_\Sigma(P_\Sigma)$. We conclude that choosing P_Σ minimizing $\beta_\Sigma(P)$ as was proposed in the noiseless case in [1] is a reasonable notion of optimal projection within the class of algorithms $\{\mathcal{I}_{GBP}(P_\Sigma, L)\}$ independently of the backprojection L. We illustrate the importance of the restricted Lipschitz condition in the case of sparse recovery in Section B in the appendix. We also immediately remark that if we want to adapt the backprojection to improve stability to noise in $\{\mathcal{I}_{GBP}(P_\Sigma, L)\}$, we will face trade-offs in terms of convergence rate and identifiability through the RIC $\delta(\mu LA)$. Consequently, in Section 4, we discuss how the generalized back-projection L can be adapted to control the stability constant. In Section 5, we show how the approximate projection error can be controlled in the training of deep projective priors with idempotent regularization.

4. Trade-offs for stable linear recovery with GPGD with generalized backprojections

In this section, to simplify the exposition and focus on the adaptation of GPGD to noise, we suppose that $\hat{x} \in \Sigma$ and that P_Σ is a restricted β-Lipschitz projection onto Σ. We discuss how the back-projection can be adapted to structured sparse noise (outliers) through the generalized back-projection L. Suppose e is a s-sparse noise with unbounded energy (i.e. $\|e\|_2$ is very large). This can model saturation noise, occlusions, or dead samples in signal and image processing, where the amplitude information is completely lost or if the sensor is saturated (e.g., in very bright light

conditions). In this case, we cannot hope to acceptably bound $\|Le\|_2$ if L is full rank (we might be able to trade off some convergence speed for improved stability constant, see Section C in the Appendix). However, in these cases, we can often estimate the support of the noise. For saturated pixels, we just need to select pixels equal to 1 for images coded in $[0,1]^N$. Then, for any $L = BS$ where S is a diagonal matrix in $\{0,1\}^{m\times m}$ that selects the complement of the support of e (precisely, $S = \mathrm{diag}(1_{\mathrm{supp}(e)^c})$) and $B \in \mathbb{R}^{N\times m}$, we have

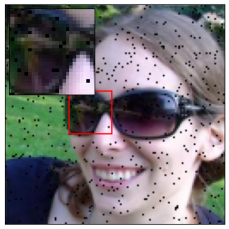
$$\|Le\|_2 = \|BSe\|_2 = 0. \tag{12}$$

Take e.g. $L = A^T S$, we get that $\delta(LA) = \delta(A^T SA)$ and $\|Le\|_2 = \|A^T Se\|_2 = \|0\|_2 = 0$, which gives

$$\|x_n - \hat{x}\|_2 \leq (\delta(A^T SA)\beta)^n \|x_0 - \hat{x}\|_2 \tag{13}$$

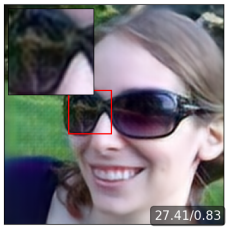
Thus, stable linear recovery is achieved if the operator $A^T SA$ has RIC with $\delta(A^T SA)\beta < 1$. This can be seen as a trade-off between stability to noise and identifiability and convergence speed as the thresholding operation removes measurements. We illustrate these results for deep projective priors (and sparse recovery in Annex D).

In Figure 1, we show that robustness to outliers is achieved by adapting the back-projection. We use the plug-and-play approach (the approximate projection P is a learned denoiser, see next sections for more details) for solving a super-resolution inverse problem for CelebA images (A is a subsampling by a factor 2). In particular, Figure 1d represents the evolution of the normalized error with respect to the number of outliers s. For different values of the number of dead samples s as represented in 1a, we solve a super-resolution inverse problem with GPGD with back-projection $A^T S$. We observe that when GPGD is adapted to this particular noise structure, it is possible to obtain a robust estimation of the ground truth which is not the case without. In addition to that, graph 1d shows that when the number of outliers increases beyond 2000 we observe a phase transition where robustness is no longer observed, a phenomenon well predicted by our theoretical findings and also observed in the case of classical sparse recovery (see Section D in the appendix).

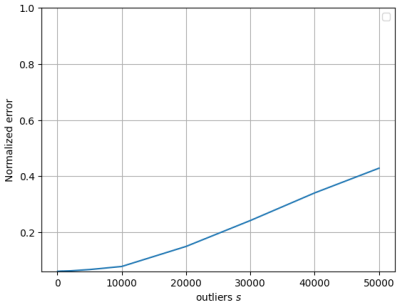


(a) Noisy image with 2000 dead pixels and a sub-sampling of factor 2

(b) GPGD recovery without adapted back-projection



(c) GPGD recovery with adapted back-projection



(d) Evolution of the normalized error for super-resolution $\times 2$ inverse problem with the number of outliers s.

Figure 1: Adaptation to sparse noise: Illustration of the trade-off between stability to noise sparsity and identifiability of Σ. We show the Normalized error bound for 90% of the experiments with respect to the sparsity s of outliers. The sparser the noise is, the greater the identifiability.

5. Mitigating the effect of approximate projections with normalized idempotent regularization

When training a projective prior P_Σ into the form of an auto-encoder or a denoiser (plug-and-play), we showed that it is sufficient to control the approximate projection constant C_{proj} to guarantee

approximate stable and robust recovery. Given X a database, deep projective priors can be trained with the following classical loss functions:

$$\begin{aligned}\mathcal{L}_{X,\text{AE}}(P) = \mathcal{L}_{X,\text{AE}}(f_D \circ f_E) &:= \sum_{x\in X} \|f_D \circ f_E(x) - x\|_2^2, \text{ for an autoencoder,} \\ \mathcal{L}_{X,\text{PnP}}(P) = \mathcal{L}_{X,\text{PnP}}(D) &:= \mathbb{E}_{x\in X,\varepsilon\sim\mathcal{N}(0,\xi^2\mathbf{I})}\left(\|D(x+\varepsilon) - x\|_2^2\right), \text{ for a denoiser.}\end{aligned} \tag{14}$$

With these loss functions, it is typically observed that $P \circ P(x) \neq P(x)$. The idempotent property that $P \circ P = P$ is a necessary condition for P to have the restricted Lipschitz property [1]. For some projective priors, this problem often leads to instabilities near convergence.

In [2], it was proposed to use a stochastic orthogonal regularization (SOR) for the training of deep projective priors to control the restricted Lipschitz constant by trying to learn an approximate orthogonal projection. SOR showed improved convergence speed and identifiability properties for very ill-posed inverse problems. However, while efficient with an oracle stopping criterion, the corresponding regularized projective priors still show some instabilities near convergence of GPGD (as shown in experiments). We propose to explore the effect of idempotent regularization of deep projective priors for the stability of GPGD. Note that this regularization has been proposed in the context of generative modeling in [22] to interpret the generative process as a projection. Our objective is to approximate an idempotent projection during training to better control the projection error term in Theorem 3.1. We define the Normalized Idempotent Regularization (NIPR) as:

$$\mathcal{R}_X(P) = \sum_{x\in X} \frac{\|P \circ P(x) - P(x)\|_2}{\|P(x)\|_2}, \tag{15}$$

and the regularized loss functions

$$\begin{aligned}\mathcal{L}^{\text{reg}}_{X,\text{AE}}(P) &= \mathcal{L}_{X,\text{AE}}(P) + \lambda\mathcal{R}_X(P); \\ \mathcal{L}^{\text{reg}}_{X,\text{PnP}}(P) &= \mathcal{L}_{X,\text{PnP}}(P) + \lambda\mathcal{R}_X(P).\end{aligned} \tag{16}$$

Note that compared to [22], we normalize the idempotent criterion to avoid a bias towards $P(x)$ of low energy. Indeed, given a linear generalized projection, consider the family of functions αP with $\alpha > 0$, we have that

$$\|(\alpha P) \circ (\alpha P)(x) - \alpha P(x)\|_2 = \alpha\|\alpha P \circ P(x) - P(x)\|_2 \to_{\alpha\to 0} 0. \tag{17}$$

Hence, this idempotent criterion without normalization can be made arbitrarily small by projections with small norm.

Experiments We now apply our NIPR to a deep denoising neural network on the CelebA dataset [26] trained with a PnP loss. We train a DRUNET denoiser [21], a U-net combined with skip connections on a dataset of size 10000. The images represent RGB faces of size 256x256. We considered two denoisers: one without NIPR and another with NIPR (with $\lambda = 0.005$). The NIPR is computed over the same batch of images as the one used from the denoising loss $\mathcal{L}_{X,\text{PnP}}$. After the training, we solve a super-resolution inverse problem by a factor of 2. We consider two image restoration algorithms: the classical unconstrained GPGD and GPGD using a regularized denoiser with NIPR. The main baseline here is PGD. Additionally, as a comparison reference, we compute the stability of the Stochastic Orthogonal Regularization (SOR [2]) added to a third DRUNET, as done for NIPR. We consider for our experiments a test set of 50 images from CelebA. To compare the recovery, we use the Peak-Signal-to-Noise-Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) of recovered images. The reader can find in the appendix additional experiments on inpainting and deblurring inverse problems. We also test our regularization over an autoencoder DPP trained on the MNIST dataset.

To study the impact of NIPR, we define two metrics that assess it. The goal in using these two stability metrics is to cover several instability cases *e.g.* to assess whether the quantity $\|x - \hat{x}\|$ is diverging or oscillating after reaching the optimal value.

The Stability Metric 1 (SM1) is defined as

$$\text{SM1}(\hat{x}, n) = \max_{i_{\min}+1 \leq i \leq i_{\min}+n} \left(\frac{\|x_i - \hat{x}\|_2}{\|x_{\min} - \hat{x}\|_2} - 1 \right), \tag{18}$$

where $\hat{x}$ is the ground truth, x_i the solution at iteration i, $x_{\min}$ the optimal one at $i_{\min}$ and n the length of the interval where we want to compute the stability. This metric captures the total deviation given a number of iterations after reaching the optimal estimate.

The Stability Metric 2 (SM2) is defined as

$$\text{SM2}(\hat{x}, n) = \sum_{i=i_{\min}+1}^{i_{\min}+n} \frac{\|x_{i+1} - x_i\|_2}{\|x_i\|_2}. \tag{19}$$

This metric captures oscillation phenomenon after reaching the optimal estimate.

Note that generally $i_{\min}$ is an oracle minimizing $\|x_i - \hat{x}\|_2$ that is not available in a real situation. Hence, controlling stability metrics can help making stopping criteria for GPGD more stable.

Table 1 represents the PSNR and the SSIM of the recovered images from a super-resolution inverse problem for the three aforementioned methods. The stability metrics SM1 and SM2 are averaged through the test set and are computed at different iterations: $n = \{i_{\min} + 10, i_{\min} + 50, i_{\min} + 100\}$, where $i_{\min}$ is the iteration of the optimal solution. Figures 2 and 3 represent the evolution of the quantity $\|x - \hat{x}\|$ for all the tests for the considered method. The graph is accompanied by visual results at different iterations: $i_{\min} + 10$, $i_{\min} + 50$, $i_{\min} + 100$. According to table 1, each method

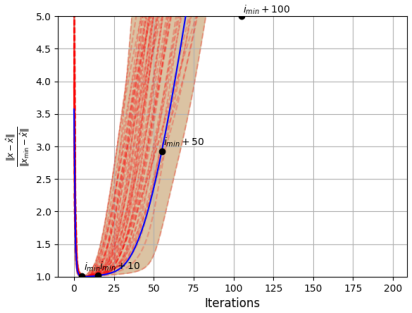
Table 1: PSNRs and stability values for a super-resolution inverse problem using GPGD without regularization, NIPR and SOR. While recovering the original image correctly, NIPR is clearly more stable after reaching $x_{\min}$ compared to PGD and SOR.

Method	PSNR↑	SSIM↑	SM1 ↓			SM2 ↓		
			$i_{\min}+10$	$i_{\min}+50$	$i_{\min}+100$	$i_{\min}+10$	$i_{\min}+50$	$i_{\min}+100$
No reg.	**34,160**	**0,933**	0,0510	3,9210	10,2517	0,018	0,257	0,799
NIPR	33,346	0,913	**0,0109**	**0,2487**	**0,7732**	**0,014**	**0,045**	**0,095**
SOR	32,855	0,913	0,0632	3,4256	18,1521	0,020	0,254	1,145

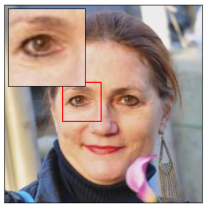
recovers the original images correctly. As it is more constrained, NIPR performs slightly worse than GPGD without regularization but still has acceptable performance. However, when comparing the stability metrics, we see clearly that NIPR has a significant impact. In fact, for both SM1 and SM2, deviations are limited for NIPR, whereas they increase significantly through the iterations for GPGD with no regularization and SOR. In particular, this result shows that NIPR has, in fact, guaranteed a more stable convergence, compared to the two other methods. It is interesting to note that even though SOR reaches the optimal solution faster, it becomes less stable compared to other methods in return. Visual results in figures 2 and 3 confirm our observations. GPGD without regularisation diverges quickly after reaching $x_{\min}$, thus producing heavily degraded images. Conversely, NIPR still produces acceptable recoveries even if we are far from the optimal solution.

6. Conclusion

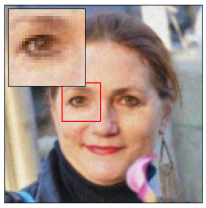
We presented an extended convergence analysis of the generalized projected gradient descent algorithm by taking into account generalized backprojections, model error and projection error. The result exposes in particular how each of these factors are affected by the restricted Lipschitz constant. We showed how we can adapt the backprojection to adapt GPGD to structured noise and what trade-offs result from this adaptation. To control the projection error, we proposed a Normalized IdemPotent Regularization (NIPR) on Deep projective priors improving experimentally the stability achieved by such GPGD methods.



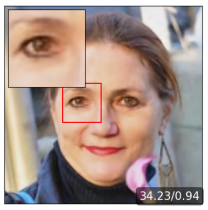
(a) Evolution of $\frac{\|x_i - \hat{x}\|}{\|x_{\min} - \hat{x}\|}$



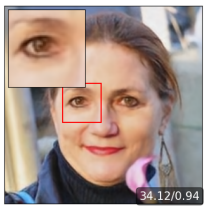
(b) Ground truth



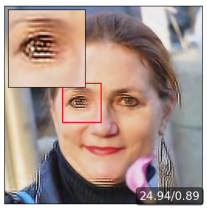
(c) Observed



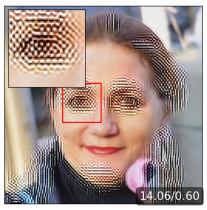
(d) x at $i_{\min}$ ($x_{\min}$)



(e) x at $i_{\min} + 10$

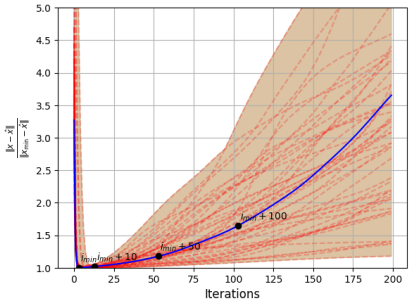


(f) x at $i_{\min} + 50$

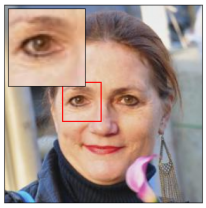


(g) x at $i_{\min} + 100$

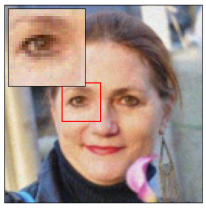
Figure 2: Super-resolution of images using GPGD without regularization. (a) represents recovery error for 50 images. The blue curve is associated with the recovery of (b). The quantity $x - \hat{x}$ quickly diverges after reaching the optimal solution and the images become unusable.



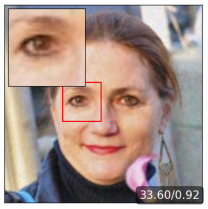
(a) Evolution of $\frac{\|x_i - \hat{x}\|}{\|x_{\min} - \hat{x}\|}$



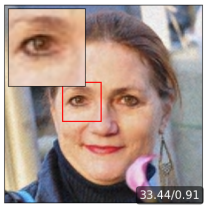
(b) Ground truth



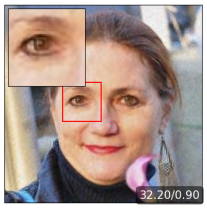
(c) Observed



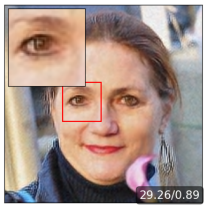
(d) x at $i_{\min}$ ($x_{\min}$)



(e) x at $i_{\min} + 10$



(f) x at $i_{\min} + 50$



(g) x at $i_{\min} + 100$

Figure 3: Super-resolution of images using NIPR. (a) represents recovery error for 50 images. The blue curve is associated with the recovery of (b). The quantity $x - \hat{x}$ is slowly diverging after reaching the optimal solution. Yet the obtained images are still recovering the original image correctly.

The results presented in this article lead to the following future works. First, automatically estimating the back-projection for different structured noises and understanding how theoretical guarantees are affected would be a natural extension of adaptation to sparse noise. Second, understanding if we could use jointly idempotent regularization, and other regularizations such as stochastic orthogonal regularization, would also be interesting.

7. Acknowledgements

Experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Universit´e de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see https://www.plafrim.fr). We are grateful to Antoine Guennec for providing us an initial code. This work was supported by PEPR PDE AI. We thank the anonymous reviewers whose comments helped improve this article.

References

[1] Yann Traonmilin, Jean François Aujol, and Antoine Guennec. Towards optimal algorithms for the recovery of low-dimensional models with linear rates. *arXiv preprint arXiv:2410.06607*, 2024.

[2] Ali Joundi, Yann Traonmilin, and Alasdair Newson. Stochastic orthogonal regularization for deep projective priors. *arXiv preprint arXiv:2505.13078*, 2025.

[3] Thomas Blumensath and Mike E Davies. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE Journal of selected topics in signal processing*, 4(2):298–309, 2010.

[4] Simon Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM Journal on numerical analysis*, 49(6):2543–2563, 2011.

[5] Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Sharp time–data tradeoffs for linear inverse problems. *IEEE Transactions on Information Theory*, 64(6):4129–4158, 2017.

[6] Mohammad Golbabaee and Mike E Davies. Inexact gradient projection and fast data driven compressed sensing. *IEEE Transactions on Information Theory*, 64(10):6707–6721, 2018.

[7] Sohail Bahmani, Petros T Boufounos, and Bhiksha Raj. Learning model-based sparsity via projected gradient descent. *IEEE Transactions on Information Theory*, 62(4):2092–2099, 2016.

[8] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

[9] Guillaume Olikier and Irène Waldspurger. Projected gradient descent accumulates at bouligand stationary points. *arXiv preprint arXiv:2403.02530*, 2024.

[10] Haoyang Liu and Rina Foygel Barber. Between hard and soft thresholding: optimal iterative thresholding algorithms. *Information and Inference: A Journal of the IMA*, 9(4):899–933, 2020.

[11] Ulugbek S Kamilov and Hassan Mansour. Learning optimal nonlinearities for iterative thresholding algorithms. *IEEE Signal Processing Letters*, 23(5):747–751, 2016.

[12] Esa Ollila, Hyon-Jung Kim, and Visa Koivunen. Robust iterative hard thresholding for compressed sensing. In *2014 6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pages 226–229. IEEE, 2014.

[13] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. *Advances in neural information processing systems*, 28, 2015.

[14] Björn Popilka, Simon Setzer, and Gabriele Steidl. Signal recovery from incomplete measurements in the presence of outliers. *Inverse Problems and Imaging*, 1(4):661–672, 2007.

[15] Christoph Studer, Patrick Kuppinger, Graeme Pope, and Helmut Bolcskei. Recovery of sparsely corrupted signals. *IEEE Transactions on Information Theory*, 58(5):3115–3130, 2011.

[16] Yann Traonmilin, Saïd Ladjal, and Andrés Almansa. Robust multi-image processing with optimal sparse regularization. *Journal of Mathematical Imaging and Vision*, 51:413–429, 2015.

[17] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE global conference on signal and information processing*, pages 945–948. IEEE, 2013.

[18] Regev Cohen, Michael Elad, and Peyman Milanfar. Regularization by denoising via fixed-point projection (red-pro). *SIAM Journal on Imaging Sciences*, 14(3):1374–1406, 2021.

[19] Wei Chen, David Wipf, and Miguel 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)*, pages 8098–8102. IEEE, 2021.

[20] Ulugbek S Kamilov, Charles A Bouman, Gregery T Buzzard, and Brendt Wohlberg. Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 40(1):85–97, 2023.

[21] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021.

[22] Assaf Shocher, Amil V Dravid, Yossi Gandelsman, Inbar Mosseri, Michael Rubinstein, and Alexei A Efros. Idempotent generative network. In *The Twelfth International Conference on Learning Representations*, 2024.

[23] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[24] Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

[25] Anthony Bourrier, Mike E Davies, Tomer Peleg, Patrick Pérez, and Rémi Gribonval. Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. *IEEE Transactions on Information Theory*, 60(12):7928–7946, 2014.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[27] Antoine Guennec, Jean-François Aujol, and Yann Traonmilin. Joint structure-texture low-dimensional modeling for image decomposition with a plug-and-play framework. *SIAM Journal on Imaging Sciences*, 18(2):1344–1371, 2025.

ç

A. Proof of Theorem 3.1

Proof of Theorem 3.1. For any n, we bound the quantity

$$\begin{aligned}
\|x_{n+1} - P_\Sigma(\hat{x})\|_2 &= \|P(x_n) - \mu L(AP(x_n) - y) - P_\Sigma(\hat{x})\|_2 \\
&= \|P(x_n) - \mu LA(P(x_n) - \hat{x}) + \mu Le - P_\Sigma(\hat{x})\|_2 \\
&= \|P(x_n) - \mu LA(P(x_n) - P_\Sigma(\hat{x}) + P_\Sigma(\hat{x}) - \hat{x}) + \mu Le - P_\Sigma(\hat{x})\|_2 \\
&= \|(I - \mu LA)(P(x_n) - P_\Sigma(\hat{x})) + \mu LA(\hat{x} - P_\Sigma(\hat{x})) + \mu Le\|_2 \\
&= \|(I - \mu LA)(P(x_n) - P_\Sigma(x_n)) \\
&\quad + (I - \mu LA)(P_\Sigma(x_n) - P_\Sigma(\hat{x})) + \mu LA(\hat{x} - P_\Sigma(\hat{x})) + \mu Le\|_2
\end{aligned} \tag{20}$$

With the triangle inequality and the RIC, we have

$$\begin{aligned}\|x_{n+1} - P_\Sigma(\hat{x})\|_2 &\leq \|(I - \mu LA)(P_\Sigma(x_n) - P_\Sigma(\hat{x}))\|_2 \\ &+ \|(I - \mu LA)(P(x_n) - P_\Sigma(x_n)) + \mu LA(\hat{x} - P_\Sigma(\hat{x})) + \mu Le\|_2 \\ &\leq \delta\|P_\Sigma(x_n) - P_\Sigma(\hat{x})\|_2 + \|(I - \mu LA)(P(x_n) - P_\Sigma(x_n))\|_2 \\ &+ \|\mu LA(P_\Sigma(\hat{x}) - \hat{x})\|_2 + \|\mu Le\|_2\end{aligned} \tag{21}$$

With the restricted β-Lipschitz condition of P_Σ (which implies $P_\Sigma(P_\Sigma(x)) = P_\Sigma(x)$, see [1]), we have

$$\begin{aligned}\|x_{n+1} - P_\Sigma(\hat{x})\|_2 \leq \delta\beta\|x_n - P_\Sigma(\hat{x})\| + \|I - \mu LA\|_{\text{op}}\|P(x_n) - P_\Sigma(x_n)\|_2 \\ + \|\mu LA\|_{\text{op}}\|P_\Sigma(\hat{x}) - \hat{x}\|_2 + \|\mu Le\|_2\end{aligned} \tag{22}$$

As $P(x_n) - P_\Sigma(x_n) = R(x_n)$, by hypothesis on R and by definition of $d_{P_\Sigma}(\hat{x}, \Sigma)$, we obtain

$$\|x_{n+1} - P_\Sigma(\hat{x})\|_2 \leq \delta\beta\|x_n - P_\Sigma(\hat{x})\| + \|I - \mu LA\|_{\text{op}}\eta + \|\mu LA\|_{\text{op}}d_{P_\Sigma}(\hat{x}, \Sigma) + \|\mu Le\|_2 \tag{23}$$

Let $\xi = \|I - \mu LA\|_{\text{op}}\eta + \|\mu LA\|_{\text{op}}d_{P_\Sigma}(\hat{x}, \Sigma) + \|\mu Le\|_2$. We show by induction

$$\|x_{n+1} - P_\Sigma(\hat{x})\|_2 \leq (\delta\beta)^{n+1}\|x_0 - P_\Sigma(\hat{x})\|_2 + \left(\sum_{i=0}^{n}(\delta\beta)^i\right)\xi \tag{24}$$

For $n = 0$, this is exactly (23).

Suppose step n true, with (23) (at step $n + 1$), we have:

$$\begin{aligned}\|x_{n+2} - P_\Sigma(\hat{x})\|_2 &\leq \delta\beta\left((\delta\beta)^{n+1}\|x_0 - P_\Sigma(\hat{x})\|_2 + \left(\sum_{i=0}^{n}(\delta\beta)^i\right)\xi\right) + \xi \\ &= (\delta\beta)^{n+2}\|x_0 - P_\Sigma(\hat{x})\|_2 + \left(\sum_{i=1}^{n+1}(\delta\beta)^i\right)\xi + \xi \\ &= (\delta\beta)^{n+2}\|x_0 - P_\Sigma(\hat{x})\|_2 + \left(\sum_{i=0}^{n+1}(\delta\beta)^i\right)\xi.\end{aligned} \tag{25}$$

This shows the induction.

We also have for any n, when $\delta\beta < 1$,

$$\begin{aligned}\|x_n - \hat{x}\|_2 &\leq \|x_n - P_\Sigma(\hat{x}) + P_\Sigma(\hat{x}) - \hat{x}\|_2 \\ &\leq \|x_n - P_\Sigma(\hat{x})\|_2 + \|P_\Sigma(\hat{x}) - \hat{x}\|_2 \\ &\leq (\delta\beta)^n\|x_0 - P_\Sigma(\hat{x})\|_2 + \left(\sum_{i=0}^{n-1}(\delta\beta)^i\right)\xi + d(\hat{x}, \Sigma) \\ &\leq (\delta\beta)^n\|x_0 - P_\Sigma(\hat{x})\|_2 + \frac{\|(I - \mu LA)\|_{\text{op}}\eta}{1 - \delta\beta} + \left(1 + \frac{\|\mu LA)\|_{\text{op}}}{1 - \delta\beta}\right)d(\hat{x}, \Sigma) \\ &+ \frac{\|\mu Le\|_2}{1 - \delta\beta}\end{aligned} \tag{26}$$

which concludes the proof. □

B. A numerical illustration of the importance of restricted Lipschitz for stable recovery

We consider the problem of sparse recovery, i.e. $\Sigma = \Sigma_k$ the set of k-sparse vectors with PGD. It has been shown that the orthogonal projection (which amounts to performing iterative hard thresholding) is restricted Lipschitz with constant $\beta = \sqrt{\frac{3+\sqrt{5}}{2}}$ which is close to optimal for restricted

Lipschitz projections onto Σ_k [1]. We propose to consider projections P_α that deteriorate the restricted Lipschitz constant of the orthogonal projection. We define, for any $z \in \mathbb{R}^N$:

$$P_\alpha(z) := \begin{cases} \left(1 + \alpha \frac{\|z - P_\Sigma^\perp(z)\|_2}{\|P_\Sigma^\perp(z)\|_2}\right) P_\Sigma^\perp(z) \text{ if } P_\Sigma^\perp(z) \neq 0. \\ 0 \text{ otherwise.} \end{cases} \tag{27}$$

where the orthogonal projection is defined in Definition 2.3 (and is the hard thresholding operator when the model is Σ_k).

Lemma B.1. *Let $\Sigma = \Sigma_k$. Consider P_α defined in (27) then*

$$\beta_\Sigma(P_\alpha) \leq \beta_\Sigma(P_\Sigma^\perp) + \alpha. \tag{28}$$

Proof of Lemma B.1. For any z, $x \in \Sigma$, we have, by definition of the restricted β-Lipschitz property (of $P_\Sigma^\perp$)

$$\begin{aligned} \|P_\alpha(z) - x\|_2 &\leq \|P_\alpha(z) - P_\Sigma^\perp(z)\|_2 + \|P_\Sigma^\perp(z) - x\|_2 \\ &\leq \alpha \| \frac{\|z - P_\Sigma^\perp(z)\|_2}{\|P_\Sigma^\perp(z)\|_2} P_\Sigma^\perp(z) \| + \beta_\Sigma(P_\Sigma^\perp) \|z - x\|_2 \\ &= \alpha \|z - P_\Sigma^\perp(z)\|_2 + \beta_\Sigma(P_\Sigma^\perp) \|z - x\|_2 \end{aligned} \tag{29}$$

By definition of $P_\Sigma^\perp(z)$, as $x \in \Sigma$, we have $\|z - P_\Sigma^\perp(z)\|_2 \leq \|z - x\|_2$ and

$$\|P_\alpha(z) - x\|_2 = (\alpha + \beta_\Sigma(P_\Sigma^\perp)) \|z - x\|_2. \tag{30}$$

□

We remark that $\beta_\Sigma(P_\alpha) \to_{\alpha \to 0} \beta_\Sigma(P_\Sigma^\perp)$, and that we control the Lipschitz constant P_α with the parameter α.

In Figure 4 (top), we perform stable sparse recovery experiments with different sparsities, with IHT (PGD with $P_0 = P_\Sigma^\perp$) and PGD with P_α ($\alpha \neq 0$). For each considered sparsity of $\hat{x}$, we perform 50 experiments and plot the normalized ℓ^2 reconstruction error (thresholded by 1) of the 95% centile. We observe that increasing α and thus degrading β diminishes the identifiability properties of PGD with P_α.

For a given sparsity, where we observe stable recovery for all α (Figure 4 (bottom)), we show the convergence of the different PGD algorithms for a fixed step μ (chosen to be the largest to obtain convergence of IHT). We observe that the convergence rate is decreased when α increases, thus matching Theorem 3.1. Note that the stability is not changed in these experiments. We attribute this to the fact that the support of $\hat{x}$ is necessarily identified for stable recovery, leading to $P_\alpha(z) \approx P_\Sigma^\perp(z)$ when z is close to Σ.

C. Impact of the step size μ on the stability constant

Let us consider the case $L = A^T$. In this case, we have $C_{\text{stab}} := \frac{\mu}{1 - \delta(\mu A^T A)\beta}$. As $\delta(\mu A^T A)$ can be interpreted as an operator norm of $I - \mu A^T A$ *restricted* to the low-dimensional model Σ, it is not clear how $\delta(\mu A^T A)$ behaves with respect to μ (except that $\lim_{\mu \to 0} \delta(\mu A^T A) = 1$ and that we do not verify convergence hypotheses for small μ). In the context of sparse recovery where PGD with the orthogonal projection is Iterative Hard Tresholding, we illustrate a trade-off between convergence speed and quality of recovery. Indeed, while we generally look a the largest possible μ for fast convergence (when interpreted as a gradient step), we observe that lowering μ can improve stability at the expense of identifiability and convergence speed. In Figure 5, we represent the recovery error

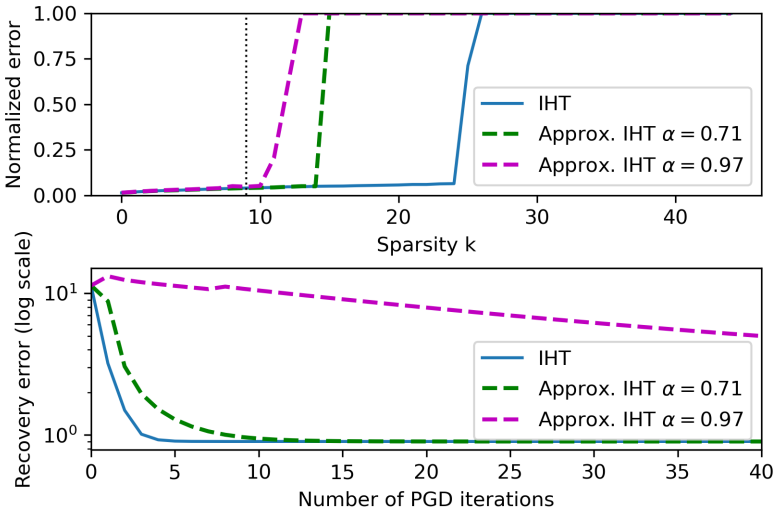


Figure 4: Importance of the restricted Lipschitz constant for stable recovery: the case of sparse recovery. Top: Normalized error bound for 95% of the experiments with respect to the sparsity k of the unknown. Bottom: convergence for one experiment with $k = 9$. We observe that worsening the Lipschitz constant (through the parameter α) deteriorates both the convergence rate and the identifiability properties of the algorithm.

with respect to sparsity of the worst 10th centile for sparse recovery ($m = 150, n = 300$, noisy random Gaussian measurements with fixed noise variance). We plot the convergence of recovery error $\|x^* - \hat{x}\|_2$ for a fixed sparsity. We observe that the best μ for stable recovery $\mu = 0.6$ allows for stable recovery of sparsities ≤ 15 while the best $\mu = 0.3$ improves noise stability for sparsity $k = 4$ at the expense of reduced identifiability and convergence speed.

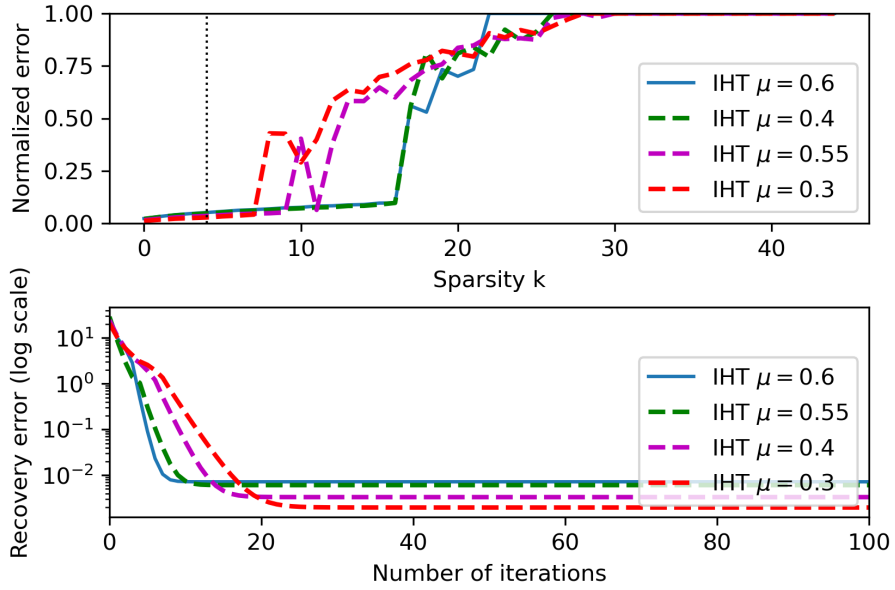


Figure 5: Impact of the step size on the stability of PGD for sparse recovery. Top: phase transition diagram for stable recovery. Bottom: impact of μ on convergence for $k = 4$

D. Trade-offs for stable linear sparse recovery with GPGD (IHT)

In the context of sparse recovery and random Gaussian measurements, a restricted isometry property of $A^T A$ is guaranteed with high probability under the condition that the number of measurements $m \geq Ck\log(n/k)$ for some potentially large constant C. As S selects $m-s$ measurements in A, we have that SA has a restricted isometry with high probability if $m \geq s + Ck\log(m/k)$. Qualitatively, fast stable recovery with iterative hard thresholding is possible if the number of measurements is $O(s+k)$ and the trade-off between identifiability of sparse vectors and robustness to sparse noise is explicit. We discuss in Section E how to adapt GPGD to unknown noise support by recasting the problem with joint models of signal and noise.

In Figure 6, we illustrate the trade-off between noise adaptation and identifiability in the case of sparse recovery with a random measurement operator. For different values of sparsity k, we add a high amplitude outlier noise of sparsity s to a low energy Gaussian noise and perform a $m-s$ hard thresholding of the residual $Ax_n - y$ in PGD (i.e. iterative hard thresholding). As predicted by the theory, the trade-off between k and s drives the success of the algorithm through the restricted isometry constant of $\mu A^T SA$ (here S is the support selected by the hard thresholding operator on the residual).

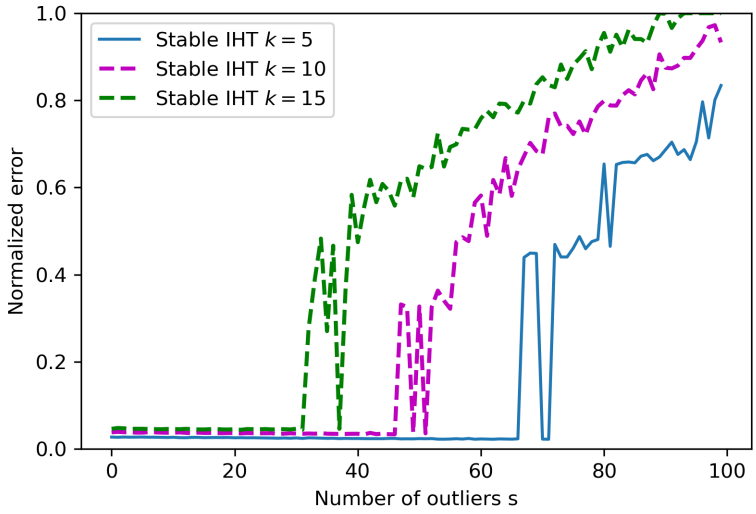


Figure 6: Adaptation to sparse noise in IHT and sparse recovery: Illustration of the trade-off between stability to noise sparsity and identifiability of sparse vectors. We show the Normalized error bound for 90% of the experiments with respect to the sparsity s of outliers for three different sparsities. The sparser the noise is, the greater the identifiability properties of stable IHT.

E. A remark on sparse noise estimation with GPGD

When the support of the sparse noise is not known, note that the problem of robust low-dimensional recovery can be recast as a simple low-dimensional recovery problem where we consider $\tilde{\Sigma} = \Sigma \times \Sigma_{noise}$ and

$$\tilde{y} = \tilde{A}\tilde{x} \tag{31}$$

where $\tilde{x} = (\hat{x}^T, e^T)^T$, $\tilde{A} = (A, I)$. The algorithm then estimates at each step both the unknown and the noise. The RIP condition notoriously relies on an incoherence between noise and sparsity model (i.e. separated support of noise and gradients for gradient-sparse image recovery [16]). It is natural to study the question of the choice of optimal $P_{\tilde{\Sigma}}$ in this case, as the low-dimensional model for signal and noise can be defined as a product model. In this case, we can construct the

optimal projection for the restricted Lipschitz constant by simply concatenating optimal projections for each model set. Recall that we can define the ℓ^2-norm in a product of Euclidean spaces $E_1 \times E_2$ by $\|(x_1, x_2)\|_2 := \sqrt{\|x_1\|_2^2 + \|x_2\|_2^2}$ for all $(x_1, x_2) \in E_1 \times E_2$. We have the following Lemma.

Lemma E.1. *Let* $\Sigma = \Sigma_1 \times \Sigma_2$, $P_i \in \arg\min_P \beta_{\Sigma_i}(P)$. *Consider* $P_\Sigma(z) = P_\Sigma(z_1, z_2) = (P_1(z_1), P_2(z_2))$. *Then* $P_\Sigma \in \arg\min_P \beta_\Sigma(P)$.

Proof of Lemma E.1. Suppose, w.l.o.g that $\beta_2 \geq \beta_1$. Let $z \in \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}, x \in \Sigma$. We have

$$\begin{aligned} \|P_\Sigma(z) - x\|_2^2 = \|P_1(z_1) - x_1\|_2^2 + \|P_2(z_2) - x_2\|_2^2 \leq \beta_1^2 \|z_1 - x_1\|_2^2 + \beta_2^2 \|z_2 - x_2\|_2^2 \\ \leq \max(\beta_1^2, \beta_2^2)(\|z_1 - x_1\|_2^2 + \|z_2 - x_2\|_2^2) = \beta_2^2 \|z - x\|_2^2 \end{aligned} \tag{32}$$

We have thus shown that $\beta_\Sigma(P_\Sigma) \leq \beta_2$.

Now let Q be a generalized projection such that $\beta_\Sigma(Q) < \beta_2$. We have, for all $z \in E$,

$$\|Q(z) - x\|_2^2 \leq \beta_\Sigma(Q)^2 \|z - x\|_2^2 \tag{33}$$

Consider $x = (x_1, x_2) \in \Sigma$, $z = (x_1, z_2) \in \Sigma_1 \times \mathbb{R}^{N_2}$. We have

$$\begin{aligned} \|Q(z) - x\|_2^2 = \|[Q(z)]_1 - x_1\|_2^2 + \|[Q(z)]_2 - x_2\|_2^2 \leq \beta_\Sigma(Q)^2 (\|x_1 - x_1\|_2^2 + \|z_2 - x_2\|_2^2) \\ = \beta_\Sigma(Q)^2 \|z_2 - x_2\|_2^2 \end{aligned} \tag{34}$$

Consider the application $\tilde{P}_2 : \mathbb{R}^{N_2} \to \mathbb{R}^{N_2}$ defined by $\tilde{P}_2(z_2) = [Q(x_1, z_2)]_2 \in \Sigma_2$. We deduce that

$$\|\tilde{P}_2(z_2) - x_2\|_2^2 = \|[Q(z)]_2 - x_2\|_2^2 \leq \beta_\Sigma(Q)^2 \|z_2 - x_2\|_2^2 \tag{35}$$

We deduce that $\beta_{\Sigma_2}(\tilde{P}_2) \leq \beta_\Sigma(Q) < \beta_2 = \beta^\star_{\Sigma_2}$, where $\beta^\star_{\Sigma_2}$ is the optimal restricted Lipschitz constant (by hypothesis), which is impossible.

We deduce that $\beta_\Sigma(Q) = \beta^\star_\Sigma$. □

We can generalize to any number of product models with the following corollary.

Corollary E.1. *Let* $\Sigma = \Sigma_1 \times \ldots \times \Sigma_q$, $P_i \in \arg\min_P \beta_{\Sigma_i}(P)$. *Consider* $P_\Sigma(z) = P_\Sigma(z) = (P_1(z_1), \ldots, P_q(z_q))$. *Then* $P_\Sigma \in \arg\min_P \beta_\Sigma(P)$.

Proof of Corollary E.1. By induction on q, for $q = 2$, use Lemma E.1.

Suppose this corollary true for some q, for $q+1$ consider $\tilde{\Sigma}_1 = \Sigma_1 \times \ldots \times \Sigma_q$ and $\tilde{\Sigma}_2 = \Sigma_{q+1}$. Apply the corollary to $\tilde{\Sigma}_1$ and Lemma E.1 to $\tilde{\Sigma}_1 \times \tilde{\Sigma}_2$. □

Note that we took the example of sparse corruptions of sparse models. In some applications, such as low-rank models and sparse noise (or sparse models and low-rank noise), restricted isometries can be obtained if there is sufficient incoherence between the sparsity model and the low-rank model. In a learning context, jointly learning projective priors for additive models has been explored in the context of structure-texture decomposition [27].

F. Additional experiments

F.1. Grid search and selection of λ for Plug-and-play setup

Regarding the choice of the regularization coefficient λ, we conducted a grid search of the optimal one. Due to computation limitations, we cannot realize a very fine analysis of the choice of this

Table 2: Comparison of metrics with respect to the values of λ for a super-resolution inverse problem. The λ that optimize the tradeoffs between PSNR and stability is $\lambda = 0.005$.

Model	Lambda	Denoising ↓	NIPR↓	PSNR ↑	SM1 $i_{min} + 10$ ↓	SM1 $i_{min} + 100$ ↓
No reg.	0	0,001571	0,0083	34,1689	0,0540	10,2001
NIPR	0,001	0,001575	0,0079	33,3761	0,0198	4,5984
	0,005	0,001590	0,0042	33,3592	0,0102	0,7625
	0,01	0,001603	0,0032	32,8168	0,0345	1,0753
	0,1	0,002076	0,0019	30,7212	0,0003	0,0003

hyperparameter. We compare in table 2 the Denoising loss, the NIPR, the PSNR and the stability metrics of a super-resolution inverse problem using differents λ.

We deduce that the optimal lambda we can choose is 0.005. We will report these values in the final version of the article.

F.2. Autoencoders

We propose in this subsection to train autoencoders over the MNIST dataset with and without NIPR. The size of the train set is 30000. To test our regularization for these, we propose to solve an inpainting inverse problem, and we display the graphs of $\frac{\|x-\hat{x}\|}{\|\hat{x}\|}$. Figure 7 shows that NIPR led to a more stable convergence as the curves are not increasing as much as for GPGD after reaching the optimal solution.

This shows in particular that, similarly to denoisers, NIPR can act on the stability of GPGD for another DPP, an autoencoder here.

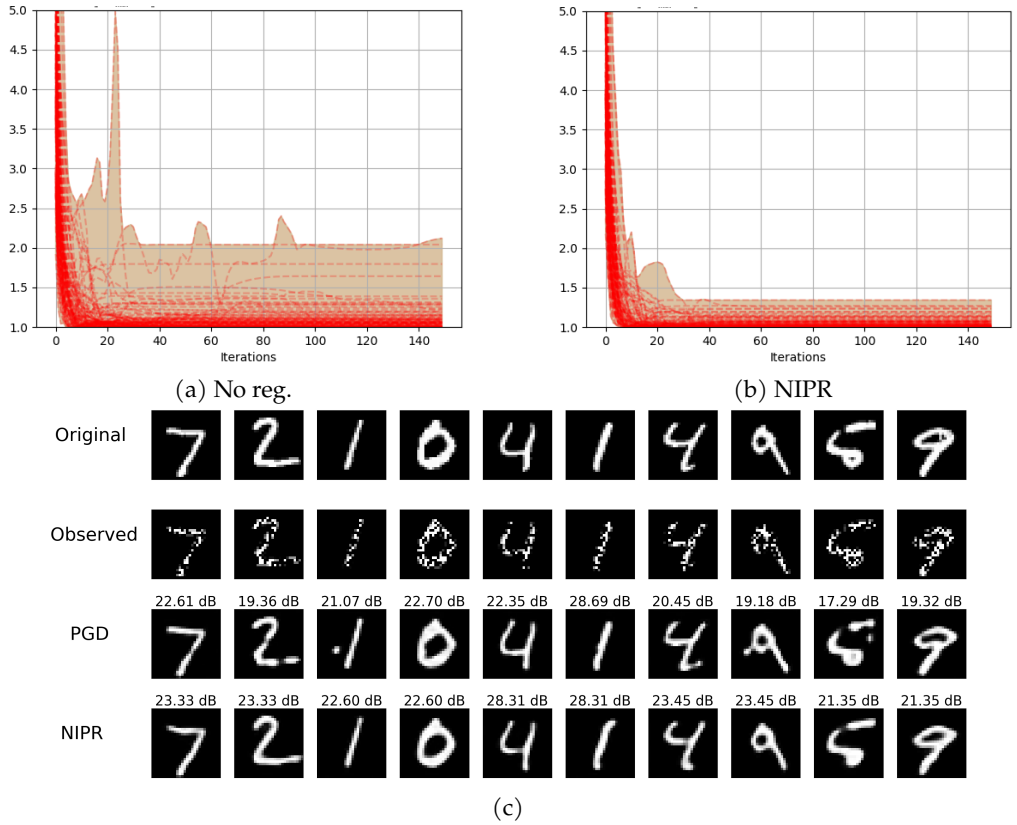


Figure 7: Inpainting of MNIST images using autoencoders with and without NIPR regularization. Once again, and for another DPP, the NIPR led to a more stable convergence compared to when we only use the vanilla GPGD algorithm.

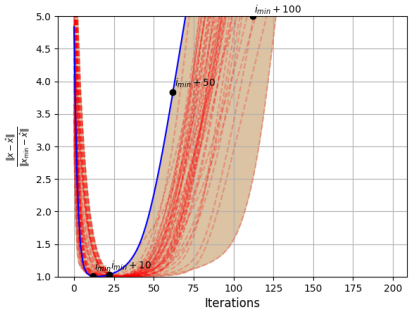
F.3. Additional experiments on denoisers

We propose in this subsection to consider additional inverse problems experiments using denoisers in their GPGD algorithms. Therefore, we consider the deblurring and inpainting tasks.

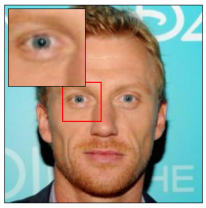
F.3.1. Deblurring

Table 3: PSNRs and stability values for a deblurring inverse problem using GPGD without regularization, NIPR and SOR. While recovering the original image correctly, NIPR is clearly more stable after reaching $x_{\min}$ compared to GPGD and SOR.

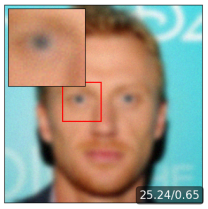
Method	PSNR↑	SSIM↑	SM1 ↓			SM2 ↓		
			$i_{\min}+10$	$i_{\min}+50$	$i_{\min}+100$	$i_{\min}+10$	$i_{\min}+50$	$i_{\min}+100$
No reg.	**29,141**	**0,832**	0,0014	1,5981	8,6641	**0,060**	0,278	1,259
NIPR	28,946	0,819	**0,0008**	**0,1795**	**0,8984**	0,068	**0,133**	**0,234**
SOR	28,147	0,797	0,0820	16,0790	37,2008	0,064	1,331	2,745



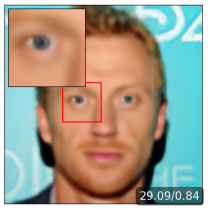
(a) Evolution of $\frac{\|x-i-\hat{x}\|_2}{\|x_{\min}-\hat{x}\|_2}$ for 50 images. The blue curve is associated to the recovery of image (b)



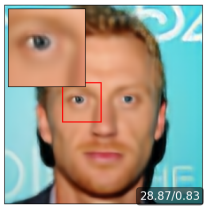
(b) Ground truth



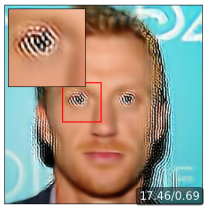
(c) Observed



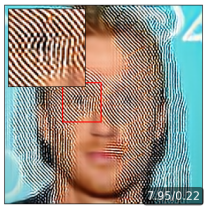
(d) x at $i_{\min}$ ($x_{\min}$)



(e) x at $i_{\min}+10$

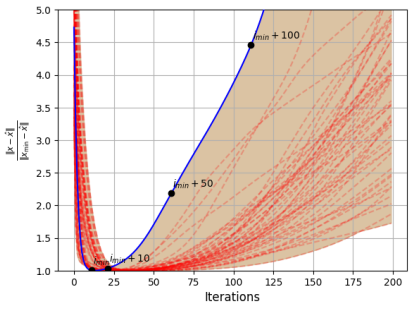


(f) x at $i_{\min}+50$

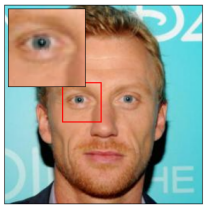


(g) x at $i_{\min}+100$

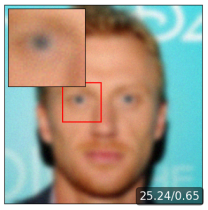
Figure 8: Deblurring of images using a GPGD algorithm without regularization. The quantity $x-\hat{x}$ quickly diverges after reaching the optimal solution and the images become unusable with several artifacts.



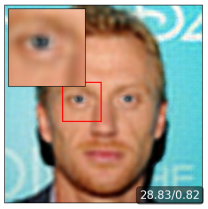
(a) Evolution of $\frac{\|x_i-\hat{x}\|_2}{\|x_{\min}-\hat{x}\|_2}$ for 50 images. The blue curve is associated to the recovery of image (b)



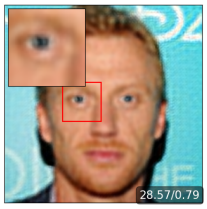
(b) Ground truth



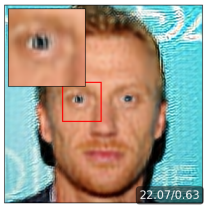
(c) Observed



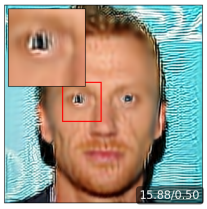
(d) x at $i_{\min}$ ($x_{\min}$)



(e) x at $i_{\min}+10$



(f) x at $i_{\min}+50$



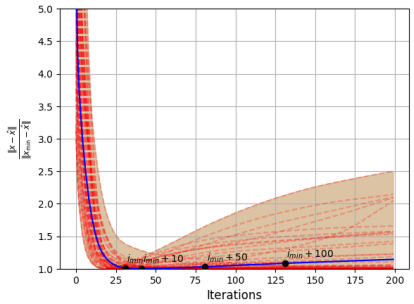
(g) x at $i_{\min}+100$

Figure 9: Deblurring of images using NIPR. The quantity $x-\hat{x}$ is diverging after reaching the optimal solution. The images degrade more slowly than GPGD without regularization.

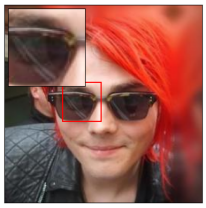
F.3.2. Inpainting

Table 4: PSNRs and stability values for an inpainting inverse problem using GPGD without regularization, NIPR and SOR. While recovering the original image correctly, NIPR is clearly more stable after reaching $x_{\min}$ compared to GPGD without regularization. For this particular inverse problem, SOR maintains a good stability.

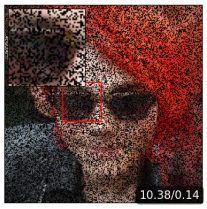
Method	PSNR↑	SSIM↑	SM1 ↓			SM2 ↓		
			$i_{\min}+10$	$i_{\min}+50$	$i_{\min}+100$	$i_{\min}+10$	$i_{\min}+50$	$i_{\min}+100$
No reg.	35,615	0,956	0,0039	0,0727	0,1376	0,088	0,468	1,197
NIPR	**36,763**	**0,959**	**0,0002**	0,0005	0,0017	**0,037**	**0,098**	0,134
SOR	34,870	0,945	**0,0002**	**0,0002**	**0,0002**	0,054	0,099	**0,113**



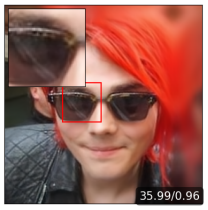
(a) Evolution of $\frac{\|x-\hat{x}\|}{\|x_{\min}-\hat{x}\|}$ for 50 images. The blue curve is associated to the recovery of image (b)



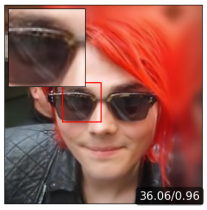
(b) Ground truth



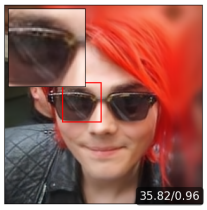
(c) Observed



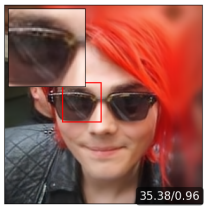
(d) x at $i_{\min}$ ($x_{\min}$)



(e) x at $i_{\min}+10$

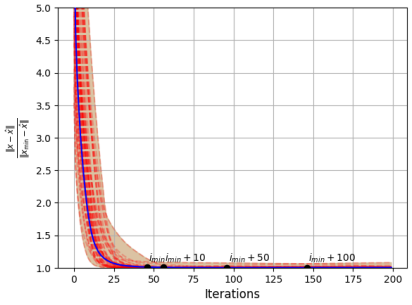


(f) x at $i_{\min}+50$

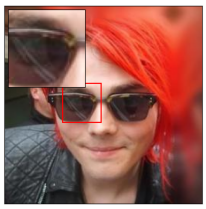


(g) x at $i_{\min}+100$

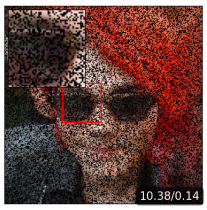
Figure 10: Inpainting of images using a PGD algorithm. The quantity $x-\hat{x}$ is slowly increasing after reaching the optimal solution but the images can still be used on average.



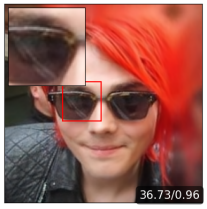
(a) Evolution of $\frac{\|x-\hat{x}\|}{\|x_{\min}-\hat{x}\|}$ for 50 images. The blue curve is associated to the recovery of image (b)



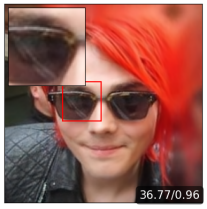
(b) Ground truth



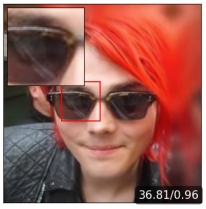
(c) Observed



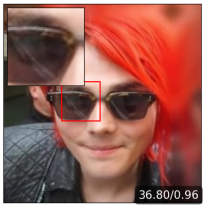
(d) x at $i_{\min}$ ($x_{\min}$)



(e) x at $i_{\min} + 10$



(f) x at $i_{\min} + 50$



(g) x at $i_{\min} + 100$

Figure 11: Inpainting of images using NIPR. The quantity $x - \hat{x}$ is converges and is stable. The images at the end of the iterations represent the original image correctly.