Auto-encodeurs atomiques parcimonie-max et application aux problèmes inverses

Ali JOUNDI[1] Yann TRAONMILIN[1] Alasdair NEWSON[2]

[1]Univ. Bordeaux, CNRS, Bordeaux INP, IMB, UMR 5251, F-33400 Talence, France.

[2]ISIR, Sorbonne Université, Paris, France.

Résumé – Un auto-encodeur atomique est une architecture de réseau de neurones (1) qui décompose une image suivant le modèle de parcimonie classique i.e. en somme d'atomes de faible dimension. Mais, celui-ci n'est pas adapté pour certains types d'images. Pour cela, on propose et on étudie un nouveau modèle de parcimonie, la *parcimonie-max*, puis on implémente l'auto-encodeur atomique associé. On montre expérimentalement qu'il conduit à une décomposition parcimonieuse des images d'entrée. Avec cette nouvelle architecture, on résout un problème inverse de super résolution via une descente de gradient projeté qui utilise le réseau entraîné comme opérateur de projection. Cette architecture montre une robustesse améliorée par rapport aux précédentes.

Abstract – An atomic autoencoder is a neural network architecture (1) that decomposes an image following the classical sparsity model i.e. as a sum of low dimensional atoms. However, it poorly represents some types of images. To this end, we propose and study the new max-sparsity model, then implement the corresponding atomic autoencoder. We show experimentally that it leads to a sparse decomposition of the input images. With this new architecture, we solve a super resolution inverse problem via a projected gradient descent that uses the trained network as a projection. This architecture shows improved robustness compared to the previous ones.

1 Introduction

Un auto-encodeur est un réseau de neurones composé d'un encodeur $f_E : \mathbb{R}^n \to \mathbb{R}^d$ projetant (par exemple une image) dans un espace latent et d'un décodeur $f_D : \mathbb{R}^d \to \mathbb{R}^n$ qui reconstruit l'entrée à partir de sa représentation latente. En d'autres termes, l'auto-encodeur f est la composition $f = f_D \circ f_E : \mathbb{R}^n \to \mathbb{R}^n$. L'entraînement de l'auto-encodeur sur une base de données X se fait en minimisant une fonction de perte $\mathcal{L}_X$ définie par :

$$\mathcal{L}_X(f) = \mathcal{L}_X(f_D \circ f_E) := \sum_{x \in X} \|f_D \circ f_E(x) - x\|_2^2. \quad (1)$$

Les auto-encodeurs ont de nombreux intérêts en traitement des images, comme la génération aléatoire de nouveaux éléments de la base de données ou la résolution de problèmes inverses mal posés. [9] donne un aperçu de plusieurs techniques sur l'utilisation de modèles génératifs dans le contexte des problèmes inverse avec quelques résultats théoriques associés. Il est possible, en particulier, d'utiliser des auto-encodeurs dans une descente de gradient projeté en tant qu'étape de projection [8]. Le succès de ces applications repose souvent sur l'accès à un espace latent structuré, ce qui reste un sujet de recherche actif dans le domaine des modèles génératifs.

Ici, on s'intéresse aux auto-encodeurs atomiques, proposés dans [7]. Ils sont étroitement liés à l'apprentissage par dictionnaire. Étant donné un dictionnaire $\mathcal{D}$ et un vecteur de poids w (souvent considéré comme parcimonieux), une image x est modélisée par $x = \mathcal{D}w$. De nombreux algorithmes ont été proposés [5] pour calculer la matrice $\mathcal{D}$, mais ils supposent un modèle linéaire entre le dictionnaire appris et les images. Ces auto-encodeurs ont donc pour but de donner une représentation parcimonieuse d'images en les décomposant en une somme d'atomes dans un dictionnaire continu que l'on implémente grâce à un réseau de neurones profond (voir Figure 1). Étant donné une image x, un auto-encodeur atomique encode x en un vecteur latent $\theta = f_E(x) \in \mathbb{R}^{kd_0}$. Le vecteur latent θ est divisé en k codes latents atomiques $\theta_i \in \mathbb{R}^{d_0}$ tels que $\theta = (\theta_i)_{i=1}^k$ qui peuvent être décodés avec un décodeur de la forme $f_D(\theta) = \sum_{i=1}^k g(\theta_i)$, où g est le décodeur atomique représentant le dictionnaire continu (l'image est la somme de k atomes paramétrés par les θ_i). Ces atomes décrivent des caractéristiques de bas niveau de l'image en entrée [7].

Ces auto-encodeurs atomiques montrent certaines limites pour fournir une représentation précise de certains types d'images. Par exemple, le modèle de sommation représente difficilement les images constituées d'objets se masquant les uns les autres (comme par exemple décrit par le modèle statistique "feuilles mortes" où chaque image est formée par la superposition de plusieurs objets indépendants et différents qui s'occultent partiellement les uns les autres [2]).

Contribution Dans cet article, on propose une amélioration des auto-encodeurs atomiques grâce à un nouveau modèle de parcimonie qui vise à mieux représenter les images naturelles en les considérant constituées d'objets qui se masquent partiellement les uns les autres.

Dans la section 2, on définit le modèle "parcimonie-max" et on présente une propriété qui le relie au modèle de parcimonie classique. On propose une modification de l'auto-encodeur atomique de [7] (SUM-AAE) pour tenir compte de ce nouveau modèle. On compare ainsi les deux auto-encodeurs atomiques et les décompositions qu'ils réalisent sur des jeux de données différents (section 3).

Enfin, dans la section 4, on utilise ces auto-encodeurs atomiques dans un algorithme de descente de gradient projeté (PGD) pour résoudre un problème inverse de super résolution et on compare les différentes reconstructions qu'ils réalisent.

2 Le modèle "parcimonie-max"

Dans l'approche "apprentissage de dictionnaire" classique, une image x est modélisée par une somme d'images élémentaires dans un dictionnaire $\mathcal{D}$ pondéré par une matrice de poids w *i.e.* $x = \mathcal{D}w$. On propose plutôt de considérer chaque image comme la superposition de différents objets indépendants qui s'occultent les uns les autres. On modélise cela via le maximum pixel par pixel d'éléments pondérés de $\mathcal{D}$ pour former la sortie. On présente ici ce modèle ainsi que son lien avec le modèle de parcimonie classique.

2.1 Definition et propriété de $\Sigma_k^{\max}$

Soit $\mathcal{D}$ un dictionnaire fini ou infini dans $\mathbb{R}^n$. Si il est fini, on définit $\mathcal{D} = \{a_i \in \mathbb{R}^n, i \leq \#\mathcal{D}\}$. $a_{i,j}$ est la j^{ieme} coordonnée de l'atome i dans $\mathcal{D}$. x_i est la composante i de $x \in \mathbb{R}^n$.

On définit le modèle de parcimonie généralisé Σ par :

$$\Sigma = \{x = \sigma(\lambda_1 a_1, \dots, \lambda_k a_k), a_i \in \mathcal{D}, \lambda_i \in \mathbb{R}\}, \quad (2)$$

où $\sigma : \mathbb{R}^{n\times k} \longrightarrow \mathbb{R}^n$ est une fonction d'agrégation et k le degré de parcimonie de l'image dans $\mathcal{D}$. Dès lors, on définit :

- Le modèle de parcimonie classique Σ_k [6] (*parcimonie-somme*) où σ est une somme, i.e. :

$$\Sigma_k := \left\{ x = \sum_{i=1}^{k} \lambda_i a_i, a_i \in \mathcal{D}, \lambda_i \in \mathbb{R} \right\}; \quad (3)$$

- le modèle *parcimonie-max* $\Sigma_k^{\max}$ où $\sigma = \text{maxel}(\cdot)$ est le maximum pixel par pixel des images i.e. :

$$\Sigma_k^{\max} := \{x = \text{maxel}(\lambda_1 a_1, \dots, \lambda_k a_k), a_i \in \mathcal{D}, \lambda_i \in \mathbb{R}\}; \quad (4)$$

où $v = \text{maxel}(u_1, \dots, u_k)$ est défini par $(v_j)_{j\leq n} = (\max(u_{1,j}, \dots, u_{k,j}))_{j\leq n}$.

Il est à noter que cette formulation générale soulève une question plus large concernant le choix de la fonction d'agrégation adéquate selon le type d'images (ou de signaux) considérées.

Dans le contexte des problèmes inverses, l'étude des propriétés structurelles de ces modèles est importante pour fournir des garanties théoriques de reconstruction (par exemple l'homogénéité a un impact dans les garanties de reconstruction [11]). Il est montré dans [3] que $\Sigma_k^{\max}$ n'est pas homogène mais qu'il peut être inclus dans le modèle de parcimonie somme sous certaines conditions :

Proposition 2.1 *Si $\mathcal{D}$ vérifie les deux conditions suivantes :*

1. *Pour tout $i \neq j$, $\text{supp}(a_i) \cap \text{supp}(a_j) = \varnothing$*
2. *Pour tout i et $j \leq n$, $k \leq n$ $a_{i,j} \cdot a_{i,k} \geq 0$*

Alors $\Sigma_k^{\max} \subsetneq \Sigma_k$. Mais en général, $\Sigma_k^{\max} \not\subset \Sigma_k$.

Ces conditions sur le dictionnaire imposent notamment un même signe pour les élements de chaque atome, ce qui rappelle la factorisation en matrices non-négatives [5]. Ainsi si les atomes sont à supports disjoints, $\Sigma_k^{\max}$ bénéficie des garanties de reconstruction de Σ_k (par simple inclusion) mais l'inverse n'est pas toujours vrai. Donc il est plus facile dans ce cas là de reconstruire les éléments $\Sigma_k^{\max}$ comparé aux éléments de Σ_k. Nos expériences montrent que l'apprentissage d'auto-encodeurs "parcimonie-max" donne des supports approximativement disjoints et donc que ces hypothèses sont raisonnables.

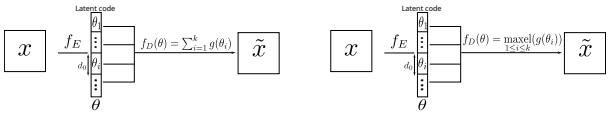
2.2 Auto-encodeurs atomiques "parcimonie-max" (MAX-AAE)

Dans [7], le modèle de parcimonie-somme est utilisé pour construire des auto-encodeurs atomiques $f = f_D \circ f_E$ tel que : $\theta = f_E(x)$ et $\tilde{x} = f_D(\theta) = \sum_{i=1}^{k} g(\theta_i)$, où f_E et g (définissant f_D) sont des réseaux de neurones (voir Figure 1). Dans cet article, on modifie le décodeur f_D pour définir un modèle de parcimonie-max.

Definition 2.1 (Auto-encodeur parcimonie-max) *Un auto-encodeur parcimonie-max est une fonction $f = f_D \circ f_E$ où $f_E : \mathbb{R}^n \to \mathbb{R}^{kd_0}$ est un encodeur et $f_D : \mathbb{R}^d \to \mathbb{R}^n$ un décodeur défini par :*

$$f_D(\theta) = \text{maxel}(g(\theta_1), \dots, g(\theta_k)), \quad (5)$$

où $g : \mathbb{R}^{d_0} \to \mathbb{R}^n$ décode un seul atome latent (voir Figure 1).



(a) SUM-AAE (b) MAX-AAE

FIGURE 1 : Architecture des auto-encodeurs parcimonie-somme et parcimonie-max

3 Décomposition via l'auto-encodeur atomique "parcimonie-max"

On compare les auto-encodeurs atomiques "parcimonie-somme" (SUM-AAE) et "parcimonie-max" (MAX-AAE) définis dans la section 2.2. Ils sont entraînés sur MNIST [1] et sur Fashion MNIST [12], qui est une base de données d'images 28x28 d'habits. L'espace latent est de taille 200 (20 blocs de taille 10). L'architecture des réseaux est présentée dans le tableau 1. Un auto-encodeur simple (AE) de même taille d'espace latent est utilisé comme référence de comparaison. Les comparaisons sont réalisées sur des bases de test de taille 600 via la métrique Peak Signal-to-Noise ratio (PSNR).

TABLE 1 : Architecture des auto-encodeurs atomique - entre parenthèses, le nombre de filtres des auto-encodeurs associés à Fashion MNIST et CIFAR10 [4].

Couches	Neurones/Filtres	Kernel	Fonction d'activation
Conv2D 1	4 (16)	3x3	LeakyReLU
Conv2D 2	8 (32)	3x3	LeakyReLU
MLP 1	210	-	LeakyReLU
Latent Vector	200	-	LeakyReLU
MLP 1	20	-	LeakyReLU
MLP 2	100	-	LeakyReLU
MLP 3	200	-	LeakyReLU
MLP 4	1568	-	LeakyReLU
Conv2D + Upsample 1	8 (32)	3x3	LeakyReLU
Conv2D + Upsample 2	16 (4)	3x3	LeakyReLU
Conv2D	1	3x3	-

Le tableau 2 donne le PSNR moyen des auto-encodages sur les bases de test. Il montre que MAX-AAE a de meilleures performances que SUM-AAE et AE sur MNIST. Cependant, pour Fashion MNIST, la performance de MAX-AAE est moindre du fait de la présence de plus d'images texturées dans cette

base de données. Néanmoins, ceci ne dégrade pas les performances dans le cadre de la super résolution puisqu'il limite les hallucinations lors de la reconstructions d'images (comparé à AE et SUM-AAE, voir section suivante).

TABLE 2 : PSNR moyens de l'auto-encodage de AE, SUM-AAE et MAX-AAE ainsi que les variances associées

	MNIST	Fashion MNIST
AE	26.72 ± 2.66 dB	$\mathbf{25.46 \pm 3.76}$ dB
SUM-AAE	27.35 ± 2.34 dB	25.28 ± 3.57 dB
MAX-AAE	$\mathbf{28.42 \pm 2.74}$ dB	24.25 ± 3.27 dB

La Figure 2 représente les décompositions de SUM-AAE et MAX-AAE sur Fashion MNIST. MAX-AAE produit une décomposition en atomes approximativement disjoints et nets (à mettre en lien avec 2.1) alors que pour SUM-AAE, les atomes sont flous et étalés. On peut vérifier qu'à la formation de l'image finale, il y a bien occlusion des atomes entre eux.

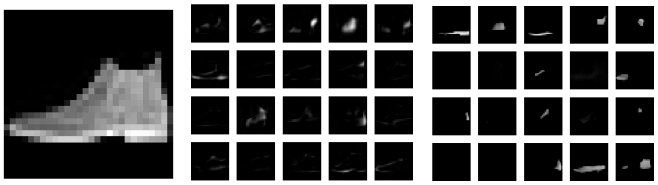


Image en entrée (a) SUM-AAE (b) MAX-AAE

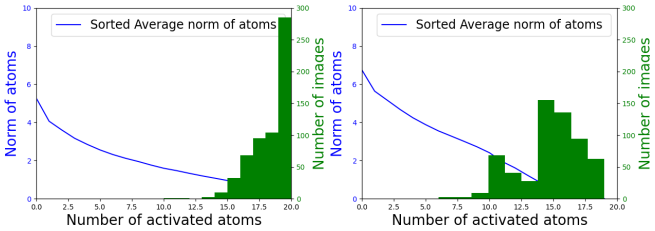
FIGURE 2 : Décomposition d'une image de Fashion MNIST via les auto-encodeurs. SUM-AAE décompose en atomes flous qui se recouvrent alors que MAX-AAE produit une décomposition parcimonieuse en atomes nets et plus disjoints.

En plus de ces décompositions améliorées, MAX-AAE produit une représentation latente plus parcimonieuse. On représente sur la Figure 3 l'histogramme du nombre d'atomes activés pour la base de données de test : un atome $a_i = g(\theta_i)$ qui forme une image x est dit activé si $\frac{\|a_i\|_2}{\|x\|_2} \geq 0.05$. On présente, de plus, la courbe moyenne de l'évolution des normes des atomes après les avoir ordonnés de manière croissante. Cette figure montre qu'en moyenne MAX-AAE active moins d'atomes que SUM-AAE : l'histogramme est plus à gauche et la norme des atomes est annulée à partir du 16ème atome pour MAX-AAE, alors qu'aucune n'est nulle pour SUM-AAE. MAX-AAE produit donc une représentation plus parcimonieuse. De plus, chaque bloc de MAX-AAE décrit une caractéristique de faible niveau de l'entrée. Par exemple Figure 2, chaque bloc activé représente une partie de la chaussure : Le premier bloc (la première image de la décomposition) représente l'arrière de la semelle.

Extension à des images naturelles : CIFAR10

On compare ici les décompositions de SUM-AAE et MAX-AAE sur une base de données d'images plus complexes. Plus précisément, on utilise CIFAR10, qui est constituée d'images de taille 32x32 réparties en 10 sous classes d'objets (avion, bateau, chien, chat etc). Les images, initialement en RGB, sont transformées en niveau de gris. On garde les mêmes architectures que précédemment. On représente Figure 4, la décomposition de deux images via SUM-AAE et MAX-AAE.

On constate que, pour SUM-AAE et MAX-AAE, la décomposition est moins parcimonieuse qu'avant (Figure 2). Ce



(a) SUM-AAE (b) MAX-AAE

FIGURE 3 : Histogramme du nombre d'atomes activés (axe de droite) et la norme moyenne des atomes en ordre décroissant (axe de gauche). L'histogramme de MAX-AAE est plus à gauche indiquant qu'en moyenne moins d'atomes sont activés.

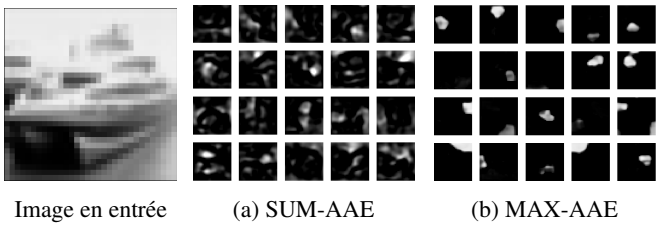


Image en entrée (a) SUM-AAE (b) MAX-AAE

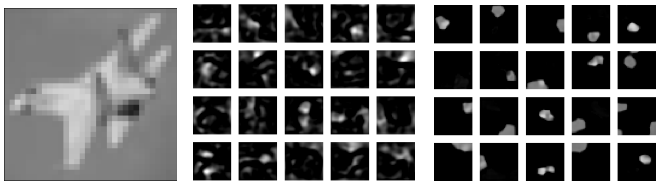


Image en entrée (c) SUM-AAE (d) MAX-AAE

FIGURE 4 : Decomposition d'images de CIFAR10 via les auto-encoders. SUM-AAE décompose en atomes flous et eparpillés alors que MAX-AAE décompose en atomes nets et localisés.

résultat peut être expliqué par la diversité de la base de données qui force les auto-encodeurs atomiques à utiliser plus de variables latentes pour représenter les caractéristiques des images. Néanmoins, les décompositions sont à nouveau différentes. MAX-AAE décompose en atomes qui représentent des zones bien délimitées de l'image alors que SUM-AAE fournit une décomposition floue et éparse. De plus, cette décomposition reste plus parcimonieuse : certains atomes ne sont pas activés pour MAX-AAE et le sont tous pour SUM-AAE.

Cette décomposition réalisée par MAX-AAE confirme donc les observations faites pour Fashion MNIST. De plus, le caractère local des atomes rappelle la segmentation d'images et suggère que MAX-AAE pourrait réaliser ces tâches.

4 Application à la super résolution

L'objectif de la super résolution d'image est d'obtenir une image de haute résolution à partir d'une image de basse résolution. Un modèle a priori de faible dimension est nécessaire pour la résolution de ce type de problème. On définit :

$$y = \mathbf{A}\hat{x}, \tag{6}$$

où $y \in \mathbb{R}^m$ est l'image sous résolue, $\hat{x} \in \mathbb{R}^n$ est l'image originale de haute résolution. Ici, $\mathbf{A} \in \mathbb{R}^{m \times n} = \mathbf{SF}$ où $\mathbf{S}$ est une matrice de sous échantillonage (d'un facteur 2 ici) et $\mathbf{F}$ est une matrice de convolution par un flou gaussien.

On cherche à résoudre :

$$\mathbf{x}^* = \underset{x \in \Sigma}{\operatorname{argmin}} \|\mathbf{A}x - y\|_2^2. \tag{7}$$

Ici, Σ est un modèle de faible dimension induit par l'auto-encodeur f. Il est donc non convexe. On utilise alors une descente de gradient projeté (PGD) définie par les itérations suivantes, étant donné une initialisation x_0 :

$$\mathbf{x}_{k+1} = P_\Sigma(x_k - \gamma \mathbf{A}^T(\mathbf{A}x_k - \mathbf{y})) \tag{8}$$

où γ est le pas et la projection P_Σ est réalisée par l'auto-encodeur appris f [8]. Pour des modèles généraux de faible dimension, cet algorithme converge linéairement vers un point fixe de la projection suivant des conditions d'isométrie restreinte sur $\mathbf{A}$ et de Lipschitz restreinte de la projection [10].

On compare les résultats de PGD avec les trois auto-encodeurs de la section 3. Le tableau 3 les compare quantitativement sur des bases de tests. Sur les deux bases de données, PGD-MAX-AAE est plus efficace que PGD-SUM-AAE et a des performances similaires à PGD-AE, et ce malgré l'architecture plus restreinte de MAX-AAE. On note à nouveau qu'à l'inverse de PGD-AE, PGD-SUM-AAE et PGD-MAX-AAE permettent la résolution du problèmes inverses tout en fournissant un accès à la décomposition de chaque image.

TABLE 3 : PSNR moyen de reconstruction. Bien que MAX-AAE soit plus contraint, PGD-MAX-AAE donne des résultats similaires à PGD-AE et meilleur que PGD-SUM-AAE.

	MNIST	Fashion MNIST
PGD-AE	26.90 ± 2.63 dB	21.95 ± 2.92 dB
PGD-SUM-AAE	25.25 ± 2.22 dB	20.91 ± 2.58 dB
PGD-MAX-AAE	$\mathbf{27.02 \pm 2.72}$ dB	$\mathbf{22.01 \pm 3.45}$ dB

On observe des différences visuelles entre PGD-MAX-AAE et PGD-AE (Figure 5). Les résultats sont similaires sur une base de donnée simple comme MNIST, mais les différences deviennent significatives quand la base de données devient plus complexe. Par exemple, dans la Figure 5, sur le t-shirt, 5ème colonne, la reconstruction fait apparaître des artefacts pour PGD-AE et PGD-SUM-AAE contrairement à PGD-MAX-AAE. On associe cela à la représentation parcimonieuse améliorée induite par le modèle parcimonie-max (voir Figure 2).

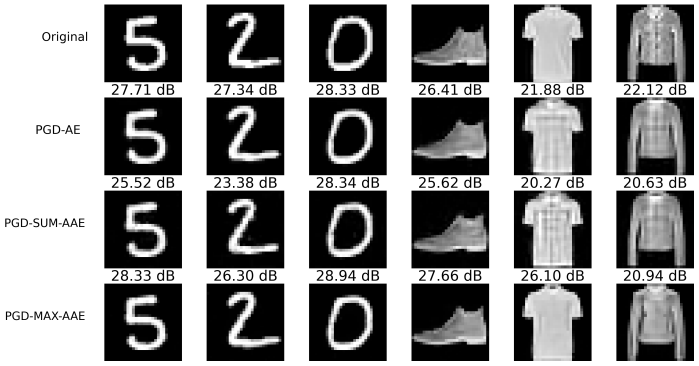


FIGURE 5 : Reconstruction de MNIST et Fashion MNIST par les PGDs avec des étapes de projection différentes. Visuellement, les performances sont similaires. Pour Fashion MNIST, PGD-MAX-AAE est plus robuste car il crée moins d'artefacts.

5 Conclusion

On a introduit un nouvel auto-encodeur atomique se basant sur le modèle de parcimonie-max. Il produit une décomposition plus nette et plus parcimonieuse de l'image en entrée.

Dans le cadre d'un problème de super résolution, cette représentation de faible dimension permet une reconstruction plus robuste comparée à celle d'un auto-encodeur simple ou celle de la proposition initiale d'auto-encodeur atomique.

La prochaine étape est d'étendre les auto-encodeurs atomiques aux images texturées photoréalistes de plus grande résolution. D'un point de vue théorique, la question du choix de la fonction d'agrégation du modèle de parcimonie selon le type d'images (ou de signaux) reste ouverte.

Références

[1] L. DENG : The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[2] Y. GOUSSEAU et F. ROUEFF : Modeling occlusion and scaling in natural images. *Multiscale Modeling & Simulation*, 6(1):105–134, 2007.

[3] A. JOUNDI, Y. TRAONMILIN et A. NEWSON : Max-sparsity atomic autoencoders with application to inverse problems. working paper or preprint, novembre 2024.

[4] A. KRIZHEVSKY, V. NAIR et G. HINTON : Cifar-10 (canadian institute for advanced research). *URL http ://www. cs. toronto. edu/kriz/cifar. html*, 5(4):1, 2010.

[5] D. LEE et S. SEUNG : Learning the parts of objects by non-negative matrix factorization. *nature*, 1999.

[6] J. MAIRAL, F. BACH et J. PONCE : Sparse modeling for image and vision processing, 2014.

[7] A. NEWSON et Y. TRAONMILIN : Disentangled latent representations of images with atomic autoencoders. *In Fourteenth International Conference on Sampling Theory and Applications*, 2023.

[8] P. PENG, S. JALALI et X. YUAN : Solving inverse problems via auto-encoders. *IEEE Journal on Selected Areas in Information Theory*, 1(1):312–323, 2020.

[9] J. SCARLETT, R. HECKEL, M. RODRIGUES, P. HAND et Y. ELDAR : Theoretical perspectives on deep learning methods in inverse problems. *IEEE journal on selected areas in information theory*, 2022.

[10] Y. TRAONMILIN, J.-F. AUJOL et A. GUENNEC : Towards optimal algorithms for the recovery of low-dimensional models with linear rates. *arXiv preprint arXiv :2410.06607*, 2024.

[11] Y. TRAONMILIN et R. GRIBONVAL : Stable recovery of low-dimensional cones in hilbert spaces : One rip to rule them all. *Applied and Computational Harmonic Analysis*, 45(1):170–205, 2018.

[12] H. XIAO, K. RASUL et R. VOLLGRAF : Fashion-mnist : a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv :1708.07747*, 2017.