Sparse overcomplete denoising: aggregation versus global optimization

Diego Carrera, Giacomo Boracchi, Alessandro Foi, and Brendt Wohlberg

***Abstract*—Denoising is often addressed via sparse coding with respect to an overcomplete dictionary. There are two main approaches when the dictionary is composed of translates of an orthonormal basis. The first, traditionally employed by techniques such as wavelet cycle-spinning, separately seeks sparsity w.r.t. each translate of the orthonormal basis, solving multiple partial optimizations and obtaining a collection of sparse approximations of the noise-free image, which are aggregated together to obtain a final estimate. The second approach, recently employed by convolutional sparse representations, instead seeks sparsity over the entire dictionary via a global optimization. It is tempting to view the former approach as providing a suboptimal solution of the latter. In this letter we analyze whether global sparsity is a desirable property, and under what conditions the global optimization provides a better solution to the denoising problem. In particular, our experimental analysis shows that the two approaches attain comparable performance in case of natural images and global optimization outperforms the simpler aggregation of partial estimates only when the image admits an extremely sparse representation. We explain this phenomenon by separately studying the bias and variance of these solutions, and by noting that the variance of the global solution increases very rapidly as the original signal becomes less and less sparse.**

***Index Terms*—Sparse Representations, Overcomplete Representations, Convolutional Sparse Coding, Denoising**

I. Introduction

Sparse representations [1] have a long and successful history in image reconstruction applications, stretching back to the classical wavelet shrinkage denoising technique [2], [3]. Denoising is often performed by computing a sparse representation of the noisy image w.r.t. an overcomplete dictionary. For dictionaries composed of all translates of an orthonormal basis, there are two main approaches. The classical approach is *cycle spinning* [3], which aggregates *partial estimates* each of which is sparse w.r.t. a different translate of the basis. An alternative is offered by *convolutional sparse representations* [4], [5], [6, Sec. II for a comprehensive review], involving a global optimization over the entire dictionary.

Intuitively, the global optimization might be expected to yield representations that are better suited for denoising. Consider the case of an image that admits a very sparse representation w.r.t. one of the orthonormal bases among the shifted copies in the dictionary: the global optimization would yield a very sparse estimate by activating only few atoms from that particular basis, while cycle spinning would aggregate all partial estimates from other shifted bases too, which might not be as sparse.

However, convolutional sparse representations have only recently begun to attract attention for solving image restoration problems [7]–[10], showing an advantage over aggregation on problems like impulse noise removal [11], but their properties are still not thoroughly understood. Surprisingly, white-noise denoising, arguably the simplest of all reconstruction problems, has been briefly mentioned in a few works addressing other issues [5], [12], [13], but has yet to receive comprehensive attention in the convolutional sparsity literature.

Our goal is to address this absence by investigating the recent convolutional sparse representations in a careful comparison against the now-classical method of wavelet cycle spinning. Our analysis is primarily meant to assess under what conditions it is more effective to solve the computationally expensive joint optimization yielding a global estimate in convolutional sparse models, rather than aggregating multiple partial estimates as in cycle spinning.

Our results show that the expected superiority of solutions from global optimization is limited to their lower bias, while their variance is often larger than that resulting from the aggregation of partial estimates. As such, global optimization outperforms the aggregation of partial estimates only when images admit an extremely sparse representation with respect to the dictionary. In contrast, when denoising natural images, the two approaches perform similarly, as the lower bias due to the global optimization is entirely offset by the larger variance.

II. Image Denoising

The input noisy image $\mathbf{s} \in \mathbb{R}^N$ corrupted by additive white Gaussian noise (AWGN) is modeled as

$$\mathbf{s} = \mathbf{y} + \boldsymbol{\eta}\,, \qquad \boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2)\,, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^N$ denotes the unknown noise-free image.

We consider denoising methods that approximate $\mathbf{y}$ as a linear combination $\widehat{\mathbf{y}}$ of atoms from a redundant set of generators of $\mathbb{R}^N$ that is formed by the union of all shifted copies $D_1, \ldots, D_N$ of a given orthonormal basis $D_1 \in \mathbb{R}^{N\times N}$:

$$\widehat{\mathbf{y}} = D\widehat{\mathbf{x}}\,, \qquad D = (\, D_1 \,\cdots\, D_N \,) \in \mathbb{R}^{N\times N^2}\,, \tag{2}$$

where $\widehat{\mathbf{x}} \in \mathbb{R}^{N^2}$ is the coefficient vector. Such redundant systems are typically used for building translation-invariant approximations of signals and images. There are two major approaches for solving (2), described below.

This work was supported by the Academy of Finland (project no. 252547) and by the U.S. Department of Energy via the LANL/LDRD Program. D. Carrera and G. Boracchi are with Politecnico di Milano, Italy (e-mail: diego.carrera@polimi.it, giacomo.boracchi@polimi.it), A. Foi is with Tampere University of Technology, Finland (e-mail: alessandro.foi@tut.fi), B. Wohlberg is with Los Alamos National Laboratory (NM), USA (e-mail: brendt@lanl.gov).

A. Aggregation of Partial Estimates

Techniques such as *cycle-spinning* [3] seek sparsity w.r.t. each orthonormal basis D_i, solving a penalized problem

$$\widehat{\mathbf{x}}_i = \arg\min_{\mathbf{u}\in\mathbb{R}^N} \frac{1}{2}\left\| D_i\mathbf{u} - \mathbf{s}\right\|_2^2 + \lambda\mathcal{R}(\mathbf{u}) \ , \quad i \in \{1,\dots,N\} \ , \tag{3}$$

where $\mathcal{R}(\cdot)$ is a regularization term promoting sparsity of $\widehat{\mathbf{x}}_i$. Since each D_i is orthonormal, problem (3) is equivalent to

$$\widehat{\mathbf{x}}_i = \arg\min_{\mathbf{u}\in\mathbb{R}^N} \frac{1}{2}\left\| \mathbf{u} - D_i^T\mathbf{s}\right\|_2^2 + \lambda\mathcal{R}(\mathbf{u}) \ , \ \ i \in \{1,\dots,N\} \ , \tag{4}$$

so that the solution $\widehat{\mathbf{x}}_i$ is given by the proximal map [14] of the regularization function $\lambda\mathcal{R}(\cdot)$. The final estimate $\widehat{\mathbf{y}}_{\mathsf{aggr}}$ is obtained aggregating the N estimates $D_i\widehat{\mathbf{x}}_i$:

$$\widehat{\mathbf{y}}_{\mathsf{aggr}} = \frac{1}{N}\sum_{i=1}^{N} D_i\widehat{\mathbf{x}}_i = D\frac{\left(\widehat{\mathbf{x}}_0^T \ \cdots \ \widehat{\mathbf{x}}_N^T\right)^T}{N} = D\widehat{\mathbf{x}}_{\mathsf{aggr}} \ . \tag{5}$$

We refer to (5) as the *aggregation* of partial estimates.

B. Global Optimization

An obvious, but more computationally expensive alternative defines a single estimate by solving a *global* optimization that jointly considers all the possible shifts of D_1 :

$$\widehat{\mathbf{x}}_{\mathsf{glob}} = \arg\min_{\mathbf{x}\in\mathbb{R}^{N^2}} \frac{1}{2}\left\| D\mathbf{x} - \mathbf{s}\right\|_2^2 + \lambda\mathcal{R}(\mathbf{x}) \ . \tag{6}$$

This problem can be formulated in an equivalent convolutional form [5], replacing $D\mathbf{x}$ by convolutions against $M \leq N$ filters:

$$\widehat{\mathbf{x}}_{\mathsf{glob}} = \arg\min_{\mathbf{x}\in\mathbb{R}^{N^2}} \frac{1}{2}\left\| \sum_{m=1}^{M} \mathbf{d}_m * \mathbf{x}_{[m]} - \mathbf{s}\right\|_2^2 + \lambda\mathcal{R}(\mathbf{x}) \ , \tag{7}$$

where $*$ denotes the convolution operator, $\mathbf{d}_m$ denotes the m^{th} column of D_1 that is used as a linear filter in the convolution, $\mathbf{x}_{[m]} \in \mathbb{R}^N$ is a subvector of $\mathbf{x}$ with $\mathbf{x}_{[m]}(j) = \mathbf{x}(m+(j-1)N)$, $j \in \{1,\dots,N\}$, and $\mathbf{x}_{[m]} \equiv \mathbf{0}$ for $m > M$. The number of filters, M, involved in the convolutional representation (7) can be smaller than N in cases where D_1 contains shifted versions of the same column, e.g., when D_1^T is a wavelet basis. In these cases, we keep only those $M \leq N$ columns of D_1 that are distinct modulo shifts, thus that correspond to different convolutional filters. The final estimate is then given by

$$\widehat{\mathbf{y}}_{\mathsf{glob}} = D\widehat{\mathbf{x}}_{\mathsf{glob}} = \sum_{m=1}^{M} \mathbf{d}_m * \widehat{\mathbf{x}}_{[m]} \ , \tag{8}$$

where the *coefficient map* $\widehat{\mathbf{x}}_{[m]}$ is the subvector gathering the representation coefficients associated with $\mathbf{d}_m$, i.e. $\widehat{\mathbf{x}}_{[m]}(j) = \widehat{\mathbf{x}}_{\mathsf{glob}}(m+(j-1)N)$, $j \in \{1,\dots,N\}$.

C. Our Analysis

Our goal is to compare these two approaches, determining whether global sparsity is a desirable property, and under what conditions the global optimization provides a better solution to the denoising problem. First, we primarily consider convex optimization problems, adopting the ℓ_1-norm as sparsity-promoting prior, for which a global minimum can be computed. Second, while the global optimization approach in the form of convolutional sparse representations has typically been applied with learned dictionaries, to fairly compare the two approaches we consider a wavelet dictionary D_1, which is fixed and not adaptively learned from training data. Third, to further investigate problems (3) and (6-7), we decompose the mean squared error (MSE) of the obtained solutions into their squared bias and variance components. The former indicates how well the approximation fits the underlying data $\mathbf{y}$ in expectation, while the latter indicates how stable this approximation is w.r.t. different realizations of the random noise $\boldsymbol{\eta}$. Finally, our analysis is performed both on natural images and synthetically generated images admitting an extremely sparse representation w.r.t. D.

III. Optimization

A. Regularization Term

To promote sparsity in the solution of (3) and (6-7), one typically adopts $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_0$ or $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_1$ as the regularization term. In these cases, the proximal maps of $\lambda\mathcal{R}$ admit closed form expressions, given by hard- and soft-thresholding [14] respectively. Since D_i is orthonormal, applying these proximal maps directly solves (3). Therefore, when $\mathcal{R}(\mathbf{x}) = \|x\|_0$, the solution of (3) is $\widehat{\mathbf{x}}_i = \mathcal{H}_\lambda(D_i^T\mathbf{s})$, where the hard-thresholding operator $\mathcal{H}_\lambda$ is defined as

$$\left[\mathcal{H}_\lambda(\mathbf{u})\right]_j = u_j \cdot 1_{\{|u_j|>\lambda\}} \ , \qquad j \in \{1,\dots,N\} \ .$$

Similarly, when $\mathcal{R}(\mathbf{x}) = \|x\|_1$, the solution of (3) is obtained as $\widehat{\mathbf{x}}_i = \mathcal{S}_\lambda(D_i^T\mathbf{s})$, where the soft-thresholding operator $\mathcal{S}_\lambda$ is

$$\left[\mathcal{S}_\lambda(\mathbf{u})\right]_j = \mathrm{sign}(u_j)\cdot\max(|u_j| - \lambda, 0) \quad j \in \{1,\dots,N\} \ .$$

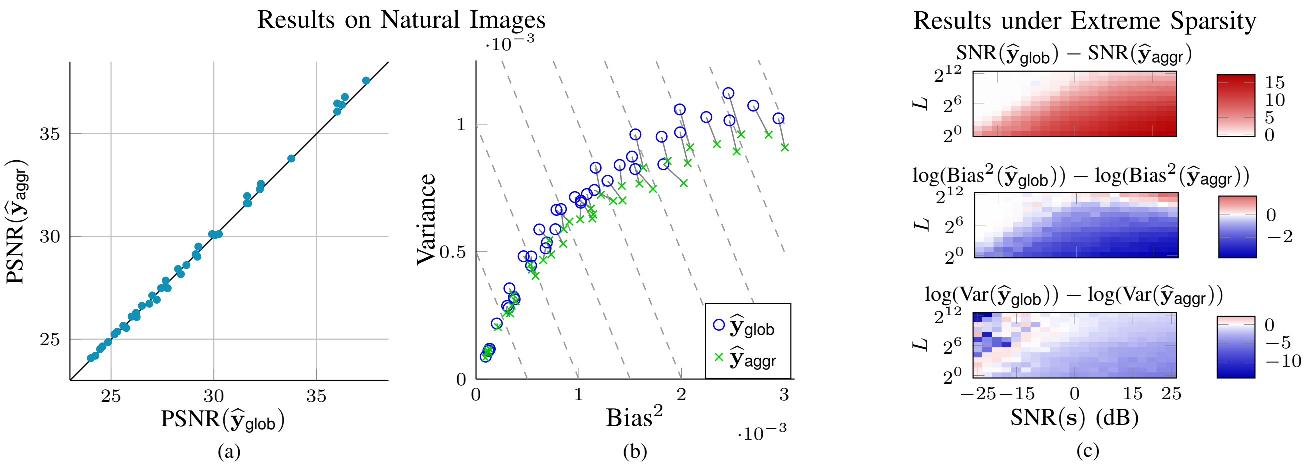
The cost of computing $\widehat{\mathbf{y}}_{\mathsf{aggr}}$ is dominated by the forward and inverse translation-invariant transform associated to D. For undecimated wavelets this is $\mathcal{O}(N\log N)$, while for a generic D with M filters it is $\mathcal{O}(MN\log N)$ through FFT.

Problem (6-7) can be approached via the Iterative Shrinkage/Thresholding Algorithm (ISTA) [15], which alternates the thresholding operator corresponding to the specific regularization term $\mathcal{R}$, with a gradient descent step on the data-fidelity term in (6-7). The cost of each iteration of ISTA is again dominated by the forward and inverse transform associated to D, needed for computing the gradient. Therefore, the cost of computing $\widehat{\mathbf{y}}_{\mathsf{glob}}$ via K ISTA iterations is K times that of aggregation. Being the convergence linear, several iterations are required to reach a sufficient approximation of the solution.

We primarily consider $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_1$ since it makes problem (6-7) convex and ISTA converges to a global minimum. In this case (6-7) can also be solved at a linear rate via a specific formulation of the Alternating Direction Method of Multipliers (ADMM) [6], whose iterations have a cost of $\mathcal{O}(MN\log N)$. When $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_0$ the problem is non-convex and ISTA is guaranteed to converge only to a critical point at a linear rate.

B. High-pass filtering

In cycle spinning, as in other wavelet approximations, coefficients from the coarsest level are not sparse [16]. Therefore,



(a) (b) (c)

Figure 1. Denoising results for $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_1$. (a) Comparison between the PSNR of $\widehat{\mathbf{y}}_{\text{glob}}$ and $\widehat{\mathbf{y}}_{\text{aggr}}$ on natural images. Equal PSNR on the diagonal. Each marker corresponds to results for a given image $\mathbf{y}$ and noise level σ. At low noise level (top-right corner) $\widehat{\mathbf{y}}_{\text{aggr}}$ outperforms $\widehat{\mathbf{y}}_{\text{glob}}$, while the two achieve similar performance as the noise gets stronger. (b) Bias-Variance decomposition of the MSE of $\widehat{\mathbf{y}}_{\text{glob}}$ and $\widehat{\mathbf{y}}_{\text{aggr}}$. Dashed anti-diagonals are the MSE level lines. The relative position of linked circles and crosses shows that although $\widehat{\mathbf{y}}_{\text{glob}}$ and $\widehat{\mathbf{y}}_{\text{aggr}}$ perform similarly, $\widehat{\mathbf{y}}_{\text{glob}}$ features lower bias and higher variance than $\widehat{\mathbf{y}}_{\text{aggr}}$. (c) Comparison on extremely sparse synthetic images. The horizontal axis reports the SNR of noisy image $\mathbf{s}$, while the vertical axis the number of nonzero coefficients L. The advantage of $\widehat{\mathbf{y}}_{\text{glob}}$ is greatest on very sparse $\mathbf{y}$ (small L) and low noise (large $\mathrm{SNR}(\mathbf{s})$). When L increases, particularly under strong noise, the relative advantage of $\widehat{\mathbf{y}}_{\text{glob}}$ degrades quickly and eventually the two estimates attain similar performance, as seen in (a).

one typically shrinks only the detail coefficients [2], [3]. This corresponds to not regularizing the approximation coefficients in (3). This is not a viable solution for the convolutional case, since if we remove the coefficient map $\mathbf{x}_{[1]}$ corresponding to the approximation filter $\mathbf{d}_1$ from $\mathcal{R}(\mathbf{x})$ in (6-7), then $\widehat{\mathbf{x}}_{[1]}$ is the deconvolution of the noisy $\mathbf{s}$ w.r.t. to $\mathbf{d}_1$. Since this solution leads to a poor estimate $\widehat{\mathbf{y}}_{\text{glob}}$, we do not perform sparse coding (6-7) directly on $\mathbf{s}$ but rather on a high-pass filtered $\mathbf{s}_h$ (as commonly done in convolutional sparse coding [11, Sec. 3]), computed by setting to 0 the approximation coefficients of the overcomplete wavelet transform D^T of $\mathbf{s}$. This is equivalent to setting $\mathbf{s}_h = \mathbf{s} - \mathbf{d}_1 * (\bar{\mathbf{d}}_1 * \mathbf{s})$, where $\bar{\mathbf{d}}_1$ denotes the conjugate filter of $\mathbf{d}_1$. Hence, we exclude $\mathbf{d}_1$ and $\mathbf{x}_{[1]}$ from the data-fidelity and regularization terms in (6-7), and add $\mathbf{s}-\mathbf{s}_h$ back to $\widehat{\mathbf{y}}_{\text{glob}}$ in (8). However, the noise affecting the high-pass filtered $\mathbf{s}_h$ is no longer white, but rather coloured by $\mathbf{v} = \boldsymbol{\delta} - \mathbf{d}_1 * \bar{\mathbf{d}}_1$, where $\boldsymbol{\delta}$ is the Dirac impulse, so the power spectrum of $\mathbf{v} * \boldsymbol{\eta}$ should be considered when denoising $\mathbf{s}_h$ in (6-7).

IV. Experiments

We perform denoising experiments on natural images as well as on synthetic data that we specifically generated to admit an extremely sparse representation w.r.t. D. We corrupt each image $\mathbf{y}$ according to (1) and compute both $\widehat{\mathbf{y}}_{\text{glob}}$ and $\widehat{\mathbf{y}}_{\text{aggr}}$. Experiments are conducted with several noise variances σ^2, and for each σ^2 we separately tune the penalty parameter λ for both methods to achieve the lowest MSE, averaged over all the considered images.

In our experiments the matrix D_1 corresponds to the orthonormal basis of the Daubechies db3 wavelet transform with 4 decomposition levels. To solve the convolutional sparse coding problem (6-7) we used the MATLAB implementation of ADMM provided in the SPORCO library [17] for $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_1$, while we relied on ISTA [15] for $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_0$.

The next two sections address $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_1$, while Section IV-C is dedicated to $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_0$.

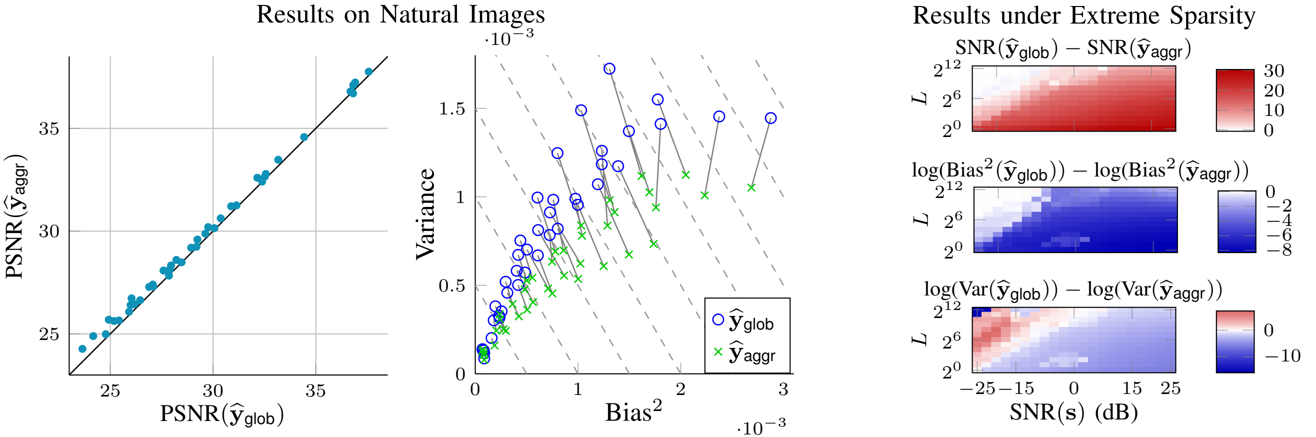
A. *Experiments on Natural Images*

We consider five test images (Lena, Barbara, Man, Peppers, Cameraman), corrupted by noise with standard deviation $\sigma \in \{5, 10, \dots, 40\}$. Each marker in Figure 1(a) represents the PSNR (average over 50 noise realizations) achieved by $\widehat{\mathbf{y}}_{\text{aggr}}$ (vertical coordinate) and $\widehat{\mathbf{y}}_{\text{glob}}$ (horizontal coordinate) for each image and σ pair. The markers are very close to the diagonal, indicating that the two methods attain very similar PSNR values, and we can see that only at low noise levels, i.e. where PSNR values are highest, the aggregation of partial estimates slightly outperforms global optimization.

In Figure 1(b) we decompose the MSE into its squared bias (horizontal coordinate) and variance (vertical coordinate) components. In these plots, anti-diagonals (dashed lines) are level lines of the MSE, and the blue circles ○ correspond to $\widehat{\mathbf{y}}_{\text{glob}}$, while the green ×-marks to $\widehat{\mathbf{y}}_{\text{aggr}}$; markers corresponding to the same pair $(\mathbf{y}, \sigma)$ are linked by a segment. The relative position of linked circles and crosses confirms that the two estimates achieve similar PSNR. Most importantly, $\widehat{\mathbf{y}}_{\text{glob}}$ features a lower bias than $\widehat{\mathbf{y}}_{\text{aggr}}$, but has a higher variance.

B. *Experiments under Extreme Sparsity*

Since the marginal performance gap between $\widehat{\mathbf{y}}_{\text{glob}}$ and $\widehat{\mathbf{y}}_{\text{aggr}}$ may appear unexpected given that global optimization should intuitively be more successful on sparse signals, we investigate how sparse the image really needs to be for our intuition to be correct, and whether the SNR plays any role in this question. We synthesize a 128×128 noise-free image $\mathbf{y} = D\mathbf{x}$ by generating $\mathbf{x}$ with L nonzero components at random positions. Then we corrupt $\mathbf{y}$ with AWGN with standard deviation σ such that the SNR of the noisy image $\mathbf{s}$ achieves a target value τ. We consider $L \in \{2^0, 2^1, \dots, 2^{12}\}$ and $\tau \in \{-25, -22.5, \dots, 25\}$, and generate 50 realizations of $\mathbf{y}$ for each pair (L, τ), and 50 realizations of $\mathbf{s}$ for each such $\mathbf{y}$.



(a) (b) (c)

Figure 2. Denoising results for $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_0$ (compare with Figure 1). (a) PSNR comparison between $\widehat{\mathbf{y}}_{\mathsf{glob}}$ and $\widehat{\mathbf{y}}_{\mathsf{aggr}}$ on natural images: $\widehat{\mathbf{y}}_{\mathsf{aggr}}$ clearly outperforms $\widehat{\mathbf{y}}_{\mathsf{glob}}$, especially under stronger noise. (b) Bias-Variance decomposition of the MSE of $\widehat{\mathbf{y}}_{\mathsf{glob}}$ and $\widehat{\mathbf{y}}_{\mathsf{aggr}}$: $\widehat{\mathbf{y}}_{\mathsf{glob}}$ suffers from a much larger variance than $\widehat{\mathbf{y}}_{\mathsf{aggr}}$. (c) Comparison on extremely sparse synthetic images: again $\widehat{\mathbf{y}}_{\mathsf{glob}}$ outperforms $\widehat{\mathbf{y}}_{\mathsf{aggr}}$ and the performance gap is much larger than for $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_1$.

Figure 1(c) shows the output SNR difference between $\widehat{\mathbf{y}}_{\mathsf{glob}}$ and $\widehat{\mathbf{y}}_{\mathsf{aggr}}$ when varying the number of nonzero coefficients L and the input SNR τ. These plots indicate that when L is small, $\widehat{\mathbf{y}}_{\mathsf{glob}}$ can achieve much larger SNR than $\widehat{\mathbf{y}}_{\mathsf{aggr}}$ thanks to its lower bias and lower variance. However, when L increases, the SNR gap shrinks and the variance of $\widehat{\mathbf{y}}_{\mathsf{glob}}$ becomes larger than that of $\widehat{\mathbf{y}}_{\mathsf{aggr}}$, especially at high noise levels. This is consistent with the results on natural images, where the two methods attain comparable PSNR and $\widehat{\mathbf{y}}_{\mathsf{glob}}$ features a larger variance. In fact, natural images arguably do not admit extremely sparse representations w.r.t. to D, and the two methods perform similar to the cases with large L in the plots of Figure 1(c).

C. Results using ℓ_0 Regularization

Figure 2 reports the denoising results for natural and for extremely sparse images using $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_0$ regularization. On natural images, $\widehat{\mathbf{y}}_{\mathsf{glob}}$ suffers from a much larger variance than $\widehat{\mathbf{y}}_{\mathsf{aggr}}$, which clearly achieves highest PSNR despite its typically higher bias. Not surprisingly, the performance gap increases with the noise level. By comparing the vertical positions of the markers in Figure 1(b) with those in Figure 2(b), we can see that the variance of $\widehat{\mathbf{y}}_{\mathsf{aggr}}$ and, especially, of $\widehat{\mathbf{y}}_{\mathsf{glob}}$ is larger when $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_0$ than when $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_1$.

The experiments on synthetic images that admit an extremely sparse representation are summarized in Figure 2(c). To deal with the lack of convexity, we initialize ISTA [15] with the extremely sparse coefficient vector $\mathbf{x}_{\mathsf{init}}$ that was used to generate $\mathbf{y}$. At least when L is small and the noise is weak, the much lower variance of $\widehat{\mathbf{y}}_{\mathsf{glob}}$ suggests that the estimate $\widehat{\mathbf{x}}_{\mathsf{glob}}$ is very close to $\mathbf{x}_{\mathsf{init}}$ and that [15] practically approaches the global minimum. Thus, on the extremely sparse images, the global optimization is confirmed to be superior to the aggregation of partial estimates also when $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_0$.

V. Discussion and Conclusions

We investigate the benefit of global optimization w.r.t. overcomplete dictionaries over aggregation of partial optimizations w.r.t. each orthogonal sub-basis, specifically comparing the convolutional sparse representations with cycle spinning. On the one hand, our experiments confirm that solving the global optimization leads to estimates that are characterized by a lower bias than the traditional aggregation of partial estimates. On the other hand, we show that the solutions of the global optimization are characterized by a larger variance, which makes the two approaches comparable when the input images are not very sparse w.r.t. the dictionary D. We speculate that the high redundancy of D, which in case of convolutional sparse representations always contains shifted atoms that are highly correlated, is the primary cause of the larger variance.

Our results indicate that solving the computationally demanding global optimization problem only has a clear advantage when D can provide a very sparse representation of the original image. When the representation is not very sparse, global optimization provides comparable performance to aggregation in the case of ℓ_1 regularization, and slightly inferior performance in the case of ℓ_0 regularization. This increased performance gap with ℓ_0 regularization highlights a practical advantage of the aggregation with orthogonal dictionaries: while switching from ℓ_1 to ℓ_0 regularization makes global optimization much more difficult, such a change does not increase the difficulty of optimizing the partial problems involving orthogonal dictionaries. Similarly, the much higher variance for the global solution on natural images when switching from ℓ_1 to ℓ_0 regularization is probably due to the non-convex nature of the optimization problem: the solutions we obtain are typically local minima, which can be expected to contribute to the overall increase in the variance.

It is unclear to us whether an adaptively learned dictionary can boost the sparsity enough to guarantee an advantage to the global optimization; the answer may be negative, as suggested by preliminary results in [13], where aggregation outperforms global optimization with learned dictionaries already at mild noise levels.

Finally, the practical advantage of aggregation can be augmented by using sparsity-adaptive weighting [18], or a recursive procedure [19], but for simplicity we aggregate with uniform weights as in classical cycle spinning.

References

[1] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Foundations and Trends in Computer Graphics and Vision*, vol. 8, no. 2-3, pp. 85–283, 2014.

[2] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[3] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising," in *Wavelets and Statistics*, A. Antoniadis and G. Oppenheim, Eds. Springer, 1995, pp. 125–150.

[4] M. S. Lewicki and T. J. Sejnowski, "Coding time-varying signals using sparse, shift-invariant representations," in *Advances in Neural Information Processing Systems 11*, 1999, pp. 730–736.

[5] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Conf. Comp. Vis. Pat. Recog. (CVPR)*, Jun. 2010, pp. 2528–2535.

[6] B. Wohlberg, "Efficient algorithms for convolutional sparse representations," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 301–315, Jan. 2016.

[7] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, "Convolutional sparse coding for image super-resolution," in *Proc. IEEE Int. Conf. Comp. Vis. (ICCV)*, Dec. 2015.

[8] T. M. Quan and W.-K. Jeong, "Compressed sensing reconstruction of dynamic contrast enhanced MRI using GPU-accelerated convolutional sparse coding," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2016, pp. 518–521.

[9] H. Zhang and V. Patel, "Convolutional sparse coding-based image decomposition," in *British Machine Vision Conference (BMVC)*, York, UK, Sep. 2016.

[10] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, 2016.

[11] B. Wohlberg, "Convolutional sparse representations as an image model for impulse noise restoration," in *Proc. IEEE Image Video Multidim. Signal Process. Workshop (IVMSP)*, Bordeaux, France, Jul. 2016.

[12] X. Luo and B. Wohlberg, "Convolutional Laplacian sparse coding," in *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, Santa Fe, NM, USA, Mar. 2016, pp. 133–136.

[13] B. Wohlberg, "Convolutional sparse representations with gradient penalties," 2017. [Online]. Available: http://arxiv.org/abs/1705.04407

[14] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[15] M. Kowalski, "Thresholding rules and iterative shrinkage/thresholding algorithm: A convergence study," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 4151–4155.

[16] I. W. Selesnick and M. A. Figueiredo, "Signal restoration with overcomplete wavelet transforms: comparison of analysis and synthesis priors," in *SPIE Optical Engineering + Applications*. International Society for Optics and Photonics, 2009, pp. 74 460D–74 460D.

[17] B. Wohlberg, "SParse Optimization Research COde (SPORCO)," Software library available from http://purl.org/brendt/software/sporco, 2016.

[18] O. G. Guleryuz, "Weighted averaging for denoising with overcomplete dictionaries," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 3020–3034, 2007.

[19] A. K. Fletcher, K. Ramchandran, and V. K. Goyal, "Wavelet denoising by recursive cycle spinning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 2, 2002, pp. 873–876.