Multi Denoising Approximate Message Passing for Optimal Recovery with Lower Computational Cost

Alessandro Perelli*, Mike E. Davies†
Institute for Digital Communications, School of Engineering, The University of Edinburgh (UK)
Email: *a.perelli@ed.ac.uk, †mike.davies@ed.ac.uk

***Abstract*—An emerging issue in large-scale inverse problems is constituted by the interdependency between computational and recovery performance; in particular in practical application, such as medical imaging, it is crucial to provide high quality estimates given bounds on computational time. While most work in this direction has gone down the lines of improving optimisation schemes, in this paper we are proposing and investigating a different approach based on a multi denoising approximate message passing (MultiD-AMP) framework for Compressive Sensing (CS) image reconstruction which exploits an hierarchy of denoisers by starting with a low fidelity model and then using the estimate as starting point for a higher fidelity models through an iterative reconstruction algorithm. MultiD-AMP achieves lower time complexity and same accuracy compared to using the same most accurate denoiser as in D-AMP. The novelty of our approach is based on exploiting the deterministic state evolution of AMP, which means the predictability of the recovery performances, to design a strategy for selecting the denoiser from a set ordered by both computational complexity and statistical efficiency. We apply the MultiD-AMP framework for image reconstruction given noisy Gaussian random linear measurements. Furthermore, we extend and show the applicability of MultiD-AMP for CS to image reconstruction.**

***Index Terms*—Compressive Sensing, Approximate Message Passing, Denoising, Computational complexity**

I. Introduction

The enormous growth in data size is becoming an issue, both for high dimensional inverse problem in medical imaging and large dataset in machine learning. Moreover, real time applications require a certain level of accuracy to be achieved within a constrained amount of time and it imposes a new perspective in designing reconstruction algorithms compared to classical iterative methods.

In machine learning problems where data is generally abundant, some works recently have considered strategies for subsampling or dynamically increase data. In particular in [1] and [2], a scheme for dynamically increasing the sample size is proposed for reducing the computational complexity of iterative methods, such as stochastic gradient descent, which generally scales with the size of training samples. For optimization and denoising problems, [3] opens up a new insight to this problem, developing the concept of "algorithmic weakening" and exploiting a hierarchy of convex relaxations, ordered by both computational and statistical efficiency. However, little has been proposed in the context of CS and other imaging problems; some recent works suggest to utilize an hierarchy of models, within proximal gradient methods, for solving composite optimization problems with lower complexity [4]. Another work [5] examined the time-data trade-off for image interpolation problem, by varying the amount of smoothing applied to the convex optimization problem. However, most of these works have addressed the problem of computational and statistical trade-off in terms of modifying or improving a statistical optimization scheme, while in this work we pursue a different approach based on building a more general family of denoisers.

In this paper we are interested in investigating and developing a computationally fast strategy for Denoising Approximate Message Passing (D-AMP) which is a compressed sensing algorithm that exhibits fast convergence and excellent performance through the application of a sequence of sophisticated denoising procedures such as BM3D [6], [7]. D-AMP can be seen as an iterative algorithm that uses sampled data to perform a sequence of updates each of them involves a gradient step, which requires the forward and back projection, and a denoising step. These denoising steps tend to dominate the computational complexity therefore, taking inspiration from [8] we propose to exploit a hierarchy of denoisers to achieve the same overall performance but at reduced computational cost.

A. Main Contribution

D-AMP algorithm [8] has been analysed in terms of inferential accuracy without considering computational complexity. This is an important missing aspect since the denoising is often the computational bottleneck in the D-AMP reconstruction. The approach that is proposed in this paper is different, we aim to derive a mechanism for leveraging a hierarchy of signal approximations and minimize the overall time complexity [3]. The intuition is based on the observation that at earlier iterations, when the estimate $\mathbf{x}^k$ is far according to some distance from the true x, the algorithm does not benefit significantly from the use of a complicated denoiser, since the attainable structure from the signal is poor. Thus, almost the same performance can be achieved through simpler faster denoisers. This leads to the idea of defining a family/hierarchy of denoisers of increased complexity; the main challenge is to define a switching scheme. This is based on an empirical finding that in MultiD-AMP we can predict exactly, in the large system limit, the evolution of the MSE. We can exploit a bound on the state evolution of a training set of images to find a proper switching strategy.

ISBN 978-0-9928626-7-1 © EURASIP 2017

II. Multi Denoising-AMP for CS system model

In this paper we consider the reconstruction of a structured signal belonging to a certain class $\mathcal{C}$, i.e. $\mathbf{x}^* \subset \mathcal{C}$, given the linear measurement

$$\mathbf{y} = \mathbf{A}\mathbf{x}^* + \mathbf{w}, \tag{1}$$

with $M < N$, where $\mathbf{y} \in \mathbb{R}^M$ is the vector of measurements, $\mathbf{A} \in \mathbb{R}^{M\times N}$ is a Gaussian random measurement matrix, $\mathbf{x}^* \in \mathbb{R}^N$ represents the signal of interest and $\mathbf{w} \in \mathbb{R}^M$ is the noise with $\mathbf{w} \sim \mathcal{N}(0, \sigma_w^2\mathbf{I})$. A denoiser is a non-linear mapping

$$D_\sigma(\cdot) : \mathbb{R}^N \to \mathbb{R}^N, \quad \mathbf{x} \to D_\sigma(\mathbf{x}) \tag{2}$$

that gives an estimate of $\mathbf{x}^*$, given some noisy measurements $\mathbf{x} = \mathbf{x}^* + \sigma\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and σ is the standard deviation of the noise. The proposed MultiD-AMP algorithm follows the D-AMP iteration:

$$\begin{aligned} \mathbf{r}^k &= \mathbf{x}^k + \mathbf{A}^T\mathbf{z}^k \\ \mathbf{x}^{k+1} &= D^k_{\sigma^k}(\mathbf{r}^k) \\ \mathbf{z}^{k+1} &= \mathbf{y} - \mathbf{A}\mathbf{x}^k + \frac{N}{M}\mathbf{z}^k \mathrm{div} D^k_{\sigma^k}(\mathbf{r}^k) \\ \sigma^{k+1} &= \frac{||\mathbf{z}^{k+1}||_2^2}{M} \end{aligned} \tag{3}$$

where $\mathbf{x}^k$ is the estimate for $\mathbf{x}^*$, $\mathbf{z}^k$ is the estimate of the residual in the measurement domain at the k-th iteration and $\mathrm{div}D^k_{\sigma^k}(\cdot)$ denotes the divergence of the denoiser. It is important to stress that the denoising functions $D^k_{\sigma^k}$ are dependent on k, i.e. for a sequence of denoisers $D^k_{\sigma^k}$, which belong to a set $\mathcal{D}$, $D^k \in \mathcal{D}$, $\forall k$.

Generally our approach can be formulated in terms of the time-data-risk class

$$\{\eta(N), M(N), R(N)\} \tag{4}$$

of estimation problems where we want to estimate a N-dimensional signal with a risk $R(N)$, given $M(N)$ measurements and with computational time $\eta(N)$.

The main challenge is to develop a switching strategy to achieve the same accuracy as the most accurate denoiser and lower computational time, given the set of denoisers $\mathcal{D}$ ordered in terms of the risk and time complexity.

The key intuition relies on exploiting a method for predicting the performance recovery for each k-th iteration, for the set of denoisers $\mathcal{D}$. This can be obtained by the State Evolution (SE) of MultiD-AMP algorithm which represents a tool to predict the expected Mean Square Error (MSE) at each iteration. While it is possible to use the MSE estimate from the SE within the D-AMP algorithm to select the best performing denoiser at each iteration, this does not account for the computation cost of each denoiser. We therefore exploit the SE given a set of trained images to determine which denoiser to select at each stage.

In the next Section, we will detail how to construct an hierarchy of denoisers based on the risk and complexity and we will develop a possible switching rule based on the deterministic SE.

III. Hierarchy of Denoisers

Before introducing the SE, we need to define the following:

Definition 1: The Risk of a denoiser $D_\sigma(\cdot)$ is defined as

$$R(\mathbf{x}^*, \sigma^2|D) = \frac{\mathbb{E}\Big[||D_\sigma(\mathbf{x}^* + \sigma\boldsymbol{\epsilon}) - \mathbf{x}^*||_2^2\Big]}{N} \tag{5}$$

where the expectation is with respect to $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$.
In this work we consider the set $\mathcal{D}$ which contains n denoisers with the property to be proper for $\mathbf{x}^* \in \mathcal{C}$, i.e.

$$\sup_{\mathbf{x}^*\in\mathcal{C}} R(\mathbf{x}^*, \sigma^2|D^k) \le \kappa\sigma^2 \tag{6}$$

and monotone

$$\forall \mathbf{x}^*, \forall \sigma_\gamma^2 \le \sigma_\xi^2,\ R(\mathbf{x}^*, \sigma_\gamma^2|D^k) \le R(\mathbf{x}^*, \sigma_\xi^2|D^k) \tag{7}$$

For this class of denoisers, we need to define an ordering for the risk and complexity. Each denoiser in the n-dimensional set $\mathcal{D}$ is indexed through 2 parameters: its risk R and the time complexity η,

$$\mathcal{D} = \{D^1_\sigma(R^1, \eta_1), \ldots, D^k_\sigma(R^k, \eta_k), \ldots, D^n_\sigma(R^n, \eta_n)\} \tag{8}$$

We propose a nested hierarchy of increasingly tighter approximations ordered over the risk

$$\sup_{\mathbf{x}^*\in\mathcal{C}} R(\mathbf{x}^*, \sigma^2|D^n) < \ldots < \sup_{\mathbf{x}^*\in\mathcal{C}} R(\mathbf{x}^*, \sigma^2|D^1) \tag{9}$$

and we expect that the runtime is inversely proportional to the risk, i.e.

$$\eta_1 < \ldots < \eta_n \tag{10}$$

The overall time of the algorithm at iteration t is

$$T(k) = \eta_{i(k)} + T(k-1) \tag{11}$$

where $k \in \mathbb{N}_+$, T is a function $T : \mathbb{N} \to \mathbb{R}$ and $\eta_{i(k)}$ denotes the i-th denoiser utilized at iteration t.
The hierarchy of denoisers for the class, defined through the relations (9) and (10), is one of the possible choices. For example, if we relax the property of the denoiser to be proper, considering the class of bounded denoisers, i.e. $||D_\sigma(\mathbf{x}) - \mathbf{x}||_2^2/N < C\sigma^2$, $\forall \mathbf{x}$, or non monotone denoisers, a different ordering in term of risk has to be defined.
Our approach aims to achieve the accuracy given by the denoiser with minimimum risk in the set, $D^n_\sigma(R^n, \eta_n)$, with a computational time $\eta < k \cdot \eta_n$, where k is the number of iterations. The strategy to determine the d optimal points

$$t^* = \{t_1, \ldots, t_d\} \tag{12}$$

across iterations where to switch denoiser is based on the prediction of the dynamics of MultiD-AMP.

ISBN 978-0-9928626-7-1 © EURASIP 2017

IV. State Evolution of MultiD-AMP

The deterministic SE generates a sequence given $\mathbf{x}^*$ as an arbitrary, fixed vector in $\mathcal{C}$

$$\begin{aligned}\theta^k(\mathbf{x}^*, \delta, \sigma_w^2, D_{\sigma^k}^k) &= \frac{\mathbb{E}\left[||D_{\sigma^k}^k(\mathbf{x}^* + \sigma^k \boldsymbol{\epsilon}) - \mathbf{x}^*||_2^2\right]}{n} \\ &= R(\mathbf{x}^*, (\sigma^k)^2, D_{\sigma^k}^k) \qquad (13)\end{aligned}$$

where $(\sigma^k)^2 = \frac{1}{\delta}\theta^k(\mathbf{x}^*, \delta, \sigma_w^2, D_{\sigma^k}^k) + \sigma_w^2$, and to note that the SE depends on $\mathbf{x}^*$. The following finding, observed in [8], shows that the SE of MultiD-AMP can predict the MSE.

Finding 1: Assuming the conditions:

- the elements $a_{ij},\ i, j \in \mathbb{R}$ of $\mathbf{A}$ are i.i.d. Gaussian, $a_{ij} \sim \mathcal{N}(0, \frac{1}{M})$;
- the noise $\mathbf{w}$ is i.i.d. Gaussian

then the State Evolution of MultiD-AMP (which is the risk of $D_{\sigma_t}^k(\cdot)$) predicts the MSE $\forall k$

$$\theta^k(\mathbf{x}^*, \delta, \sigma_w^2, D_{\sigma^k}^k) = \lim_{N \to \infty} \frac{1}{N} \|\mathbf{x}^k - \mathbf{x}^*\|_2^2 \qquad (14)$$

where we highlighted the dependency of θ^k on the true signal and the undersampling ratio $\delta = \frac{M}{N}$. Then, for finite δ and $n \to \infty$, the state evolution predicts the MSE of MultiD-AMP To get back to the class of problem (4), from Eq. (14) we obtain that the estimation error is connected, in the large system limit, with the risk and, from the relation (10), the complexity is given by the runtime.

V. Switching Strategy

At this point we are ready to define the switching strategy g as a function of the risk $R(\mathbf{x}^*, \sigma^2 | D^k)$ and the mapping from continuous to discrete time $T(k)$,

$$g(\theta(\mathbf{x}^*, \delta, \sigma_w^2, D_{\sigma^k}^k), T(k)) \qquad (15)$$

Since the SE depends on $\mathbf{x}^*$ (true signal), $\theta(\mathbf{x}^*, \delta, \sigma_w^2, D_{\sigma^k}^k)$ should be calculated a priori based on a set of "representative signals" as training data, i.e. signals $\{\mathbf{x}_{1t}, \ldots, \mathbf{x}_{Pt}\}$, where P is the dimension of the set of training signals, which belong to the same class of $\mathbf{x}^*$,

$$\{\mathbf{x}_{1t}, \ldots, \mathbf{x}_{Pt}\}, \mathbf{x}^* \in \mathcal{C} \qquad (16)$$

Given a training set of data, we generate the State Evolution and for each $T(k)$ (time T continuous variable at iteration k discrete variable); since the function θ is defined over the discrete set $T(k)$, it is meaningful to define the difference quotient function g over discrete points in the following way:

$$g(\theta^{T(k)}(\mathbf{x}^*, D_{\sigma^{T(k)}}^{T(k)}), T(k)) = \frac{\theta^{T(k)}(\cdot) - \theta^{T(k-1)}(\cdot)}{T(k) - T(k-1)} \qquad (17)$$

which can be geometrically interpreted as the angular coefficient of the straight line intersecting the 2 points.

The chosen greedy criteria for selecting the sequence t^* for switching $D_\sigma(\cdot)$ is

$$|g(\theta^{T(k+1)}(\cdot), T(k+1))| \leq \left| \frac{g(\theta^{T(k)}(\cdot), T(k))}{4} \right| \qquad (18)$$

The geometrical interpretation for the switching rule (18) is shown in Fig. 1; for each interval $[T(k-1), T(k)]$, the absolute value of the function (17) represents the slope of the tangent line to the function θ at a point $T(k_a)$, $T(k-1) < T(k_a) < T(k)$. Therefore the relation (18) indicates that we greedly decide to change denoiser when the rate of decrease of θ is less than $\frac{1}{4}$ of the previous interval.

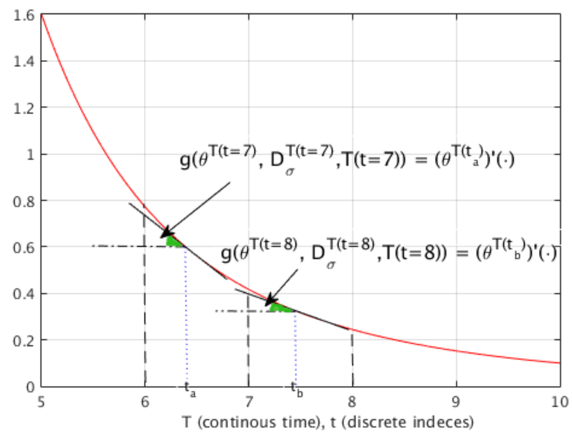

Fig. 1. Geometrical interpretation of the switching rule.

The intuition comes from analysing the behaviour of the derivative of the risk (SE) over time (complexity) evaluated at discrete time instants. The reason for using this condition is that when the slope at the discrete time $T(t+1)$ tends to decrease significantly from the one computed at $T(k)$ this implies that the rate of convergence of the "low complex" denoiser is decreasing, i.e. the denoising algorithm tends to approach the convergence and we need to switch to the denoiser with lower risk at the price of higher time complexity.

VI. Results

MultiD-AMP has been tested on 256×256 Lena image with i.i.d. Gaussian random measurements, Gaussian noise and 0.2 undersampling ratio. As a proof of principle, we use 2 denoisers, discrete Wavelet soft thresholding (DWT) and BM3D and we assume that

$$\begin{aligned}\forall \sigma,\ R(\mathbf{x}^*, \sigma^2 | D_{DWT}) &\leq R(\mathbf{x}^*, \sigma^2 | D_{BM3D}) \\ \eta_{DWT} &\leq \eta_{BM3D} \qquad (19)\end{aligned}$$

and we seek for the optimal iterate t^* to switch between the 2 denoisers, i.e. we learn one parameter t^* based on a small set of training images.

In the experiment 4 "traditional" images, boat, barbara, house, peppers have been used as training images; the switching point t^* is obtained with the greedy procedure (18) on the SE and the t^* mostly selected in the SE of each training image is the one used in the reconstruction.

Figure 3 shows that by changing the denoisers in t^*, we obtain the same MSE at convergence as the one achieved by the most powerful denoiser, i.e. BM3D, with lower computational time compared to a single denoising AMP (DWT-AMP and BM3D-AMP).

ISBN 978-0-9928626-7-1 © EURASIP 2017

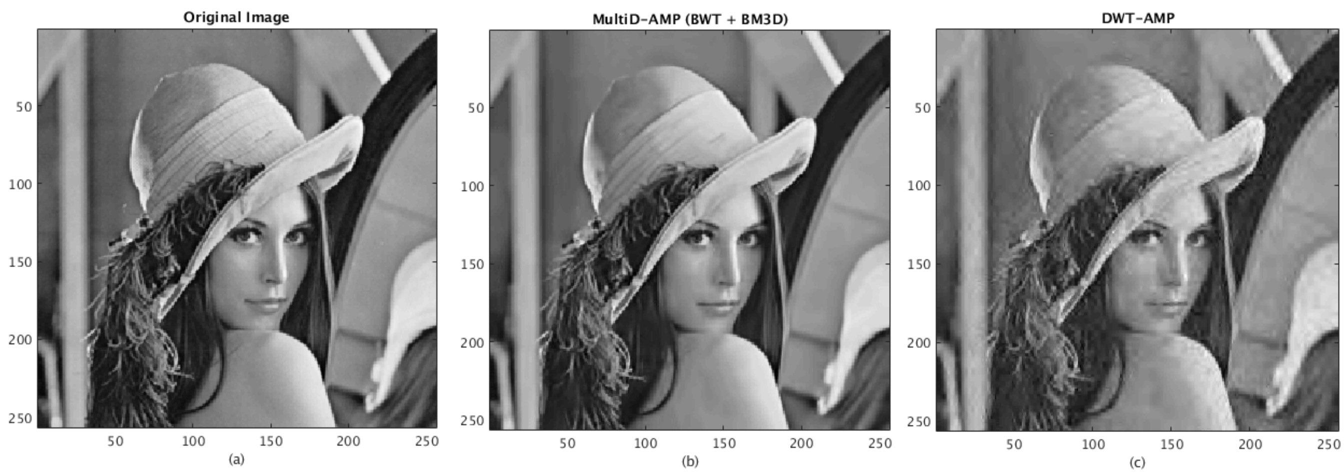

Fig. 2. CS reconstruction problem: (a) Original image, (b) MultiD-AMP, (c) DWT-AMP

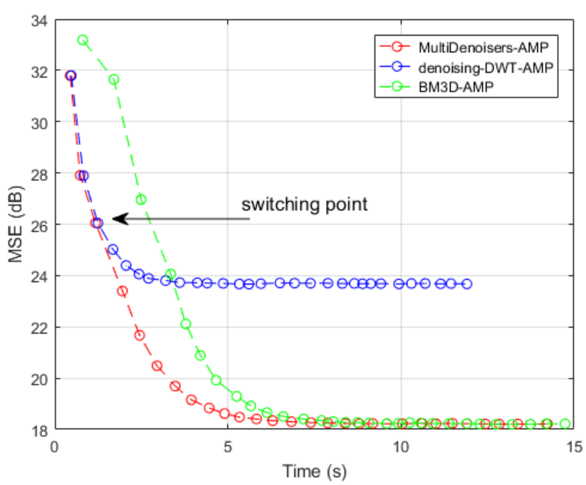

Fig. 3. MSE - time reconstruction plot: comparison MultiD-AMP, DWT-AMP, BM3D-AMP

It is worth highlighting how our strategy yields to exploit the low complexity of DWT at early iterations, when the MSE error is generally high, and then using the high accurate BM3D denoiser at the cost of higher computational time.

Additionally the proposed MultiD-AMP framework has been applied using as hierarchy of denoisers the Fast Multi-level Wavelet structure described in [9]. The CS system model is described by the algebraic relation in Eq. (1) and $\mathbf{x}$ is represented through the synthesis wavelet basis $\mathbf{W}$, $\mathbf{x} = \mathbf{W}\xi$. We consider the MultiD-AMP algorihtm described by the following update rule

$$
\begin{aligned}
\mathbf{r}^k &= \mathbf{x}^k + \mathbf{A}^T\mathbf{z}^k \\
\mathbf{x}^{k+1} &= (\mathbf{W}\mathcal{T}_{\sigma^k}\mathbf{W}^T)_{ks}(\mathbf{r}^k) \\
\mathbf{z}^{k+1} &= \mathbf{y} - \mathbf{A}\mathbf{x}^k + \frac{N}{M}\mathbf{z}^k \mathrm{div}[\mathbf{W}\mathcal{T}_{\sigma^k}\mathbf{W}^T]_{ks}(\mathbf{r}^k) \\
\sigma^{k+1} &= \frac{||\mathbf{z}^{k+1}||_2^2}{M}
\end{aligned}
$$

where $(\mathbf{W}\mathcal{T}_{\sigma^k}\mathbf{W}^T)_{ks}$ is dependent on the wavelet sub-bands s, which are changing across iterations k, by the Multi-level V-cycle scheme described in [9] and $\mathcal{T}_{\sigma^k}$ is the soft thresholding function defined for $\mathbf{x} \in \mathcal{C}$ as

$$\mathcal{T}_\eta(\mathbf{x}) = \mathrm{sgn}(\mathbf{x}) \max(|\mathbf{x}| - \eta, 0) \tag{20}$$

The Multi-level Wavelet AMP has been tested with the same setup used for the previous simulation; Fig. 4 shows the performance of the MSE in actual time for the DWT-AMP and the Multi-level Wavelet - AMP (both with soft thresholding non linear function). We can notice how exploiting a multigrid procedure in the denoising step, i.e. considering a scheme for selecting sub-bands, yields to a reduction in time complexity especially at earlier iterations; the blue curve correspond to the blue curve in Fig. 3(DWT) and as expected both the Multi-level and DWT achieve the same MSE at convergence.

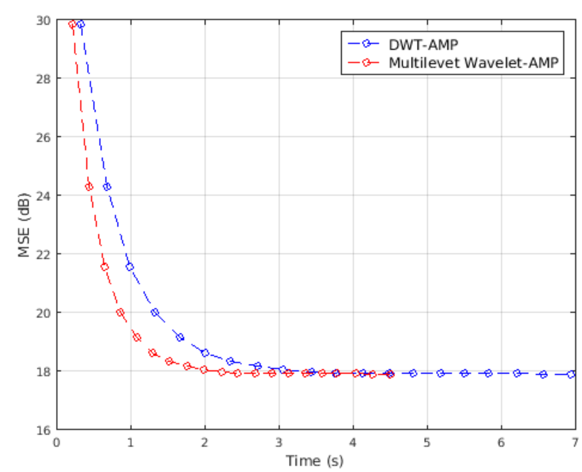

Fig. 4. MSE - time reconstruction plot: comparison Multi-level Wavelet-AMP, DWT-AMP

ISBN 978-0-9928626-7-1 © EURASIP 2017

VII. Conclusions

In this paper we have proposed a new approach in terms of risk-time trade-offs for reduce the time complexity of the family of D-AMP iterative reconstruction algorithms.

In particular a strategy for designing an hierarchy of denoisers in MultiD-AMP has been proposed and successfully tested. It is important to highlight that this framework allows the dynamic switching between denoisers based on the State Evolution which provide a powerful way to predict the performances of MultiD-AMP. Furthermore, we proposed a framework following the idea of [3] for image reconstruction and we have built a ordered set of denoisers which constitutes a more general scenario than convex nested set.

Finally we have applied the proposed MultiD-AMP concept with a Multi-level Wavelet denoiser [9] which yields to an improvement in time complexity compared to the full DWT.

Acknowledgment

The authors would like to acknowledge the support from ERC Advanced grant, project 694888, C-SENSE.

References

[1] H. Daneshmand, A. Lucchi, and T. Hofmann, “Starting small-learning with adaptive sample sizes,” in *Proceedings of the 33nd International Conference on Machine Learning, ICML*, 2016, pp. 1463–1471.

[2] T. Chen, A. Mokhtari, X. Wang, A. Ribeiro, and G. B. Giannakis, “Stochastic averaging for constrained optimization with application to online resource allocation,” *arXiv preprint arXiv:1610.02143*, 2016.

[3] V. Chandrasekaran and M. I. Jordan, “Computational and statistical tradeoffs via convex relaxation,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 13, pp. E1181–E1190, 2013.

[4] V. Hovhannisyan, P. Parpas, and S. Zafeiriou, “MAGMA: Multilevel accelerated gradient mirror descent algorithm for large-scale convex composite minimization,” *SIAM Journal on Imaging Sciences*, vol. 9, no. 4, pp. 1829–1857, 2016.

[5] J. J. Bruer, J. A. Tropp, V. Cevher, and S. R. Becker, “Designing statistical estimators that balance sample size, risk, and computational cost,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 612–624, 2015.

[6] A. Danielyan, V. Katkovnik, and K. Egiazarian, “BM3D frames and variational image deblurring,” *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1715–1728, 2012.

[7] A. Perelli and M. E. Davies, “Compressive computed tomography image reconstruction with denoising message passing algorithms,” in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 2806–2810.

[8] C. A. Metzler, A. Maleki, and R. G. Baraniuk, “From denoising to compressed sensing,” *IEEE Transactions on Information Theory*, vol. 62, no. 9, pp. 5117–5144, 2016.

[9] C. Vonesch and M. Unser, “A fast multilevel algorithm for wavelet-regularized image restoration,” *IEEE Transactions on Image Processing*, vol. 18, no. 3, pp. 509–523, 2009.