Methods and Applications of Analysis.
Vol. 26, No. 3, pp. 217–234, September 2019

A DOUGLAS-RACHFORD METHOD FOR SPARSE EXTREME LEARNING MACHINE*

TOMMI KÄRKKÄINEN† AND ROLAND GLOWINSKI‡

Abstract. Operator-splitting methods have gained popularity in various areas of computational sciences, including machine learning. In this article, we present a novel nonsmooth and nonconvex formulation and its efficient associated solution algorithm to derive a sparse predictive machine learning model. The model structure is based on the so-called extreme learning machine with randomly generated basis. Our computational experiments confirm the efficiency of the proposed method, when a bold selection of the timestep is made. Comparative tests also indicate interesting results concerning the use of the l_0 seminorm for ultimate sparsity.

Key words. Operator-splitting, Douglas-Rachford, extreme learning machine, sparse regularization.

Mathematics Subject Classification. 90C26.

1. Introduction. Operator-splitting methods have a long and versatile history in the field of computing [29]. Their origin lies in the solution of boundary value problems [28, 25, 22, 32], but more recently their role as general algorithmic frameworks for nonsmooth optimization problems resulting, e.g., from image processing [53, 46, 15, 59, 45] has increased. Along this line of enlarging interest in these methods is also Boyd et al. [8], who advanced the adoption of operator-splitting methods in machine learning, for finite dimensional problems. However, nonconvex problems were not particularly addressed in this survey. Interestingly, alternation of two basic steps to solve a nonsmooth optimization problem is also present in clustering algorithms [33].

Peaceman-Rachford [47] and Douglas-Rachford (DR) [19] methods are among classical operator-splitting approaches (see [31], Sections 2.4–2.5). Their convergence and rate or convergence for a special case of the sum of two discrete operators $A(u)+B$, a monotone and a coercive one, was analyzed in [27]. The convergence analysis concluded the dependence of the best timestep Δt on the eigenvalues of the coercive operator B, but the numerical experiments were concluded as follows: "DR performed much better, similarly as IE [Implicit Euler], in the examples where the operator A was defined in lower order spaces (by means of regularity, i.e., differentiability) than B". The work in this article builds on these preliminaries, and suggests to use Douglas-Rachford operator-splitting method for an optimization problem in machine learning having some similarity with the operators just discussed. To use DR follows also the suggestion as given in Remark 13 in [31].

The Douglas-Rachford method has been analyzed and used in the context of various problems. General convergence results for the sum of two monotone operators were provided in [41], [20], and [51]. The inverse problem related to signal recovery, modeled as the minimization of the sum of two lower semicontinuous convex functions, was addressed in [16]. DR for image denoising with multiplicative Gamma noise and bounded variation regularization, with a proof of convergence, was considered in [50].

*Received December 21, 2017; accepted for publication April 12, 2019.

†University of Jyväskylä, Faculty of Information Technology, FIN-40014, Finland (tommi.karkkainen@jyu.fi).

‡Department of Mathematics, University of Houston, Houston, TX 77204-3008, USA (roland@math.uh.edu).

In [49], interpretation of the alternating split Bregman method for image processing as a DR splitting algorithm was provided. More recently, He and Yuan [34] proved a worst-case convergence rate of DR, measured by the iteration complexity, for the sum of two maximal monotone set-valued operators. Mathematical study on the effects of the order of the two operators in DR was given in [3].

Sparsity of models is a useful target in many applications, in an attempt to use as simple model, with sufficient prediction accuracy, as possible ("occam's razor", see [7]). Often this has meant the use of l_1-norm, whose purpose is to prefer a small set of active, nonzero model parameters. A thorough treatments of the basis pursuit (BP) techniques, i.e., sparse recovery of the most important coefficients in an overcomplete collection of parameterized waveforms using the minimal l_1 norm, was given in [14]. Soft-thresholding based projection algorithms for image restoration with the l_1-norm were suggested in [5]. The fast iterative shrinkage-thresholding algorithm (FISTA) for solving sparse linear inverse problems efficiently, by using shrinkage of the gradient-descent step, was then suggested in [4]. Note that if the Lipschitz constant of the smooth, linear operator is not known, the algorihm needs to be safeguarded with backtracking to ensure descend and global convergence. Zhang et al. [56] provided convergence analysis of two general primal-dual algorithm variants with Bregman iteration for a class of problems including the l_1 regularization term. Fadili and Starck [21] suggested a Douglas-Rachford/OS method for a linear inverse problem with nonsmooth l_1-regularization. In [55], the l_1-norm was used, together with an alternating direction method with two shrinkage operators, for signal reconstruction. Repeated simpler substeps for the l_1-regularization of a linear problem were also used in [54].

Direct measurement of nonzero components, in order to prefer minimal number of them, would mean the use of the l_0-seminorm. Towards this direction, Chartrand [10] suggested to use a nonconvex regularization of the form $\int |\nabla u|^p,\ 0 < p < 1$, in image reconstruction. In connection with the soft-thresholding techniques for the l_1-norm, iterative hard-thresholding based on the l_0 seminorm was suggested in [6]. The development of such techniques (see, e.g., [58, 17, 2]) and a proximal hard-thresholding method are given in [57]. Central to our work was [11], where a maximally sparse low rank matrix approximation was obtained with l_0 regularization. For an alternating direction method, the corresponding shrinkage operator was given analytically in [11], and then generalized in [52, 13, 12]. A general summary of alternating direction methods and sparse models in given in [30], Chapters 1, 7–9, and 13. To this end, the necessity or superiority of l_0 compared to l_1 was questioned by Donoho [18], who showed that in a special linear setting of underdetermined problems, the sparsest solution can be readily obtained with the l_1-norm regularization.

The remainder of this paper, after this general introduction to the field, is as follows: Section 2 introduces the basic machine learning method, the extreme learning machine, used in the work. There, also the actual alternating direction method to obtain a sparse model is presented. The results of computational experiments with real world data sets are given in Section 3. Conclusions and final remarks are then presented in Section 4.

2. Methods. In this section, we introduce compactly the basic machine learning technique and the basic variants of the Douglas-Rachford methods.

2.1. Extreme Learning Machine. Extreme Learning Machine (ELM) is a scalable machine learning technique with randomly generated nonlinear basis. The form of basis, or kernel, has a central role in machine learning and in predictive mod-

els. Introduction of the Radial Basis Function networks (e.g., [48]) showed that the universal approximation property of neural network, i.e. capability to approximate unknown nonlinear functions similarly to the classical result of the density of polynomials, does not need a fully adaptable basis. This was the main ingredient of the Multilayered Perceptrons (see, e.g. [37, 39]), which more recently have evolved into a large pool of transformation layers of different type in deep learning (see [1] and articles therein).

ELM, with the basic methodology as described in [36, 35], is currently one of the key randomized neural network frameworks without kernel adaptation [26]. Universal approximation properties in a probabilistic sense for ELM were presented in [42, 40]. There, the need of the repeated sampling of the random kernel and the advantage of the Tikhonov regularization (weight decay in neural network lingo) were shown. The appeal of the basic ELM lies in the fact that, similarly to many other basically linear techniques [24], a regularized least-squares problem can be solved to recover the matrix of weights representing the combination of the generated, random kernel.

A predictive model is based on a given set of input-output samples representing the behavior of the unknown function. Hence, let $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^{n_0}$ and $\mathbf{y}_i \in \mathbb{R}^n$, be this data referred as the training data. To normalize the range of the initial transformation, we min-max scale $\{\mathbf{x}_i\}$ into the range $[0, 1]$. In ELM, we associate for each bias-enlarged input (to shift the random kernel from the origin) $\tilde{\mathbf{x}}_i = [1\ \mathbf{x}_i^T]^T \in \mathbb{R}^{n_0+1}$ the sigmoidal basis function $\mathbf{h}_i = \frac{1}{1+\exp(-\mathbf{G}\tilde{\mathbf{x}}_i)}$, where $\mathbf{G} \in \mathbb{R}^{m\times(n_0+1)}$ with $(\mathbf{G})_{ij} \in \mathcal{U}([-1, 1])$ (uniform distribution on $[-1, 1]$). Here m denotes the number of basis functions.

To determine the linear combination of the nonlinear generated kernel for function approximation, let us consider the discrete optimization problem

$$\min_{\mathbf{V}\in\mathbb{R}^{n\times m}} \mathcal{J}(\mathbf{V}),$$

where

$$\mathcal{J}(\mathbf{V}) = f(\mathbf{V}) + r(\mathbf{V}) = \frac{1}{2N}\sum_{i=1}^{N} \|\mathbf{V}\mathbf{h}_i - \mathbf{y}_i\|_2^2 + \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{\alpha}{2}|\mathbf{V}_{ij}|^2 + \beta|\mathbf{V}_{ij}|\right). \quad (1)$$

Notable point in the formulation above, compared to the usual discrete optimization problems, is to have the unknown in the form of a matrix and not as a vector. From the application point of view, to determine weights of the linear combination of a nonlinear basis functions, this is a natural starting point. As will be seen, this choice carries over the whole set of necessary formulations. Actually matrix would also be the natural data structure, e.g., in image processing [59, 45].

With the Fröbenius product : the variational form of (1) reads as

$$D\mathcal{J}(\mathbf{W}) : \mathbf{V} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{W}\mathbf{h}_i - \mathbf{y}_i)\cdot\mathbf{V}\mathbf{h}_i + \frac{1}{m}\left(\alpha\mathbf{W} : \mathbf{V} + \beta\partial|\mathbf{W}_{ij}| : \mathbf{V}\right) = 0 \ \ (\text{or } \ni \mathbf{0}).$$

Hence, the optimality condition for the solution $\mathbf{W} \in \mathbb{R}^{n\times m}$ of (1) reads as

$$\frac{1}{N}(\mathbf{W}\mathbf{H} - \mathbf{Y})\mathbf{H}^T + \frac{\alpha}{m}\mathbf{W} + \frac{\beta}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\partial|\mathbf{W}_{ij}| = \mathbf{0}, \quad (2)$$

where $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^N \in \mathbb{R}^{m\times N}$ and $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N \in \mathbb{R}^{n\times N}$ ($\mathbf{h}_i$'s and $\mathbf{y}_i$'s as columns).

2.2. Douglas-Rachford methods. In relation to (2), let us define

$$A_1(\mathbf{W}) = \frac{\beta}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\partial|\mathbf{W}_{ij}| \quad \text{and} \quad A_2(\mathbf{W}) = \frac{1}{N}(\mathbf{W}\mathbf{H}-\mathbf{Y})\mathbf{H}^T + \frac{\alpha}{m}\mathbf{W}.$$

Consider then the following discrete time-dependent problem for $\Delta t > 0$:

$$\frac{\mathbf{W}^{k+1}-\mathbf{W}^k}{\Delta t} + A_1(\mathbf{W}^{k+1}) + A_2(\mathbf{W}^{k+1}) = 0. \tag{3}$$

Apparently, if $\mathbf{W}^{k+1} \to \mathbf{W}^k$ as $k \to \infty$, we recover a solution satisfying (2).

For $\mathbf{W}^0 \in \mathbb{R}^{n\times m}$ given, the Douglas-Rachford operator-splitting scheme reads as follows

1) Solve $\hat{\mathbf{W}}^{k+1} \in \mathbb{R}^{n\times m}$ from

$$\frac{\hat{\mathbf{W}}^{k+1}-\mathbf{W}^k}{\Delta t} + A_1(\hat{\mathbf{W}}^{k+1}) + A_2(\mathbf{W}^k) = 0.$$

Using the well-known explicit formula for the shrinkage operator this reduces, when defining $\Gamma^k = \mathbf{W}^k - \Delta t A_2(\mathbf{W}^k)$ and $\tilde{\beta} = \frac{\beta\Delta t}{m}$, to

$$\hat{\mathbf{W}}^{k+1} = \operatorname{sgn}(\Gamma^k)\max(0, |\Gamma^k| - \tilde{\beta}), \tag{4}$$

where sgn and max are applied componentwise.

2) Define $A_1(\hat{\mathbf{W}}^{k+1}) = -\left(\frac{\hat{\mathbf{W}}^{k+1}-\mathbf{W}^k}{\Delta t} + A_2(\mathbf{W}^k)\right)$ and solve $\mathbf{W}^{k+1} \in \mathbb{R}^{n\times m}$ from

$$\frac{\mathbf{W}^{k+1}-\mathbf{W}^k}{\Delta t} + A_1(\hat{\mathbf{W}}^{k+1}) + A_2(\mathbf{W}^{k+1}) = 0.$$

Hence, "right inverse" the following identity for $\mathbf{W}^{k+1}$:

$$\mathbf{W}^{k+1}\left((1+\frac{\alpha\Delta t}{m})\mathbb{I}_{m,m} + \Delta t\mathbf{H}\mathbf{H}^T/N\right) = \mathbf{W}^n - \Delta t A_1(\hat{\mathbf{W}}^{k+1}) + \mathbf{Y}\mathbf{H}^T/N.$$

We assume that the size of the problem or the number of basis functions m is such that the system above can be solved with standard techniques. This basically restrictive assumption still allows one to address many applications of practical size, for example, in classification (see Section 3.1 and Table **??**).

To this end, let us consider replacing the nonsmooth regularization $|\mathbf{V}_{ij}|$ with the function $\frac{1}{s}|\mathbf{V}_{ij}|^s$ for $1 \gg s \to 0$. As illustrated in Figure 1, this nonconvex function approaches in the limit case $s = 0$ the seminorm l_0 measuring the cardinality of the kernel, i.e., the number of nonzero (active) basis functions used by ELM. For this purpose, we modify the regularization part of the cost function (1) as follows:

$$r(\mathbf{V}) = \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{\alpha}{2}|\mathbf{V}_{ij}|^2 + \frac{\beta}{s}|\mathbf{V}_{ij}|^s\right).$$

To cast the Douglas-Rachford method for the non-convex case, we can simply modify the Step 1. above by using the componentwise *s-shrinkage operation* [11, 13, 12]

$$\hat{\mathbf{W}}^{k+1} = \operatorname{sgn}(\Gamma^k)\max(0, |\Gamma^k| - \tilde{\beta}|\Gamma^k|^{s-1}). \tag{5}$$

Clearly, for $s = 1$, (5) coincides with (4).

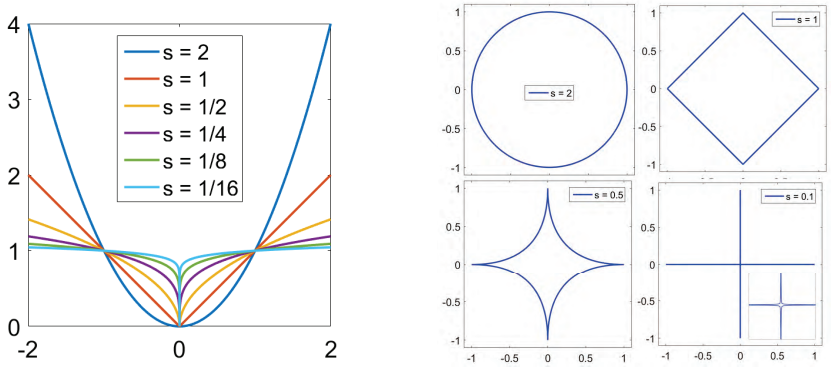


Fig. 1: *Illustrations of the l_s-norms for different values of s in 1D (left) and 2D (right).*

3. Computational experiments.

3.1. Settings. Reference versions of the techniques from the previous section were implemented with Matlab (R2015b). The computational experiments were performed with an ordinary laptop computer (64-bit operating system, 2.9 GHz processor, and 32 GB RAM). Datasets for the tests mostly originate from the UCI machine learning repository [23], for more thorough depiction see[1]. We used many of the datasets from [43, 44] because of their direct availability in a suitable format. All of the experimental datasets are described in Table **??**. There, N denotes the number of training and NT the test observations, n_0 is the size of input vectors and n defines the number of classes. The output vectors $\mathbf{y}_i$ were formed with the 1-of-n encoding, i.e., using the standard basis in $\mathbb{R}^n$ of the n classes [38]. As preprocessing, we removed the constant variables (affected variables of MNIST dataset) and, as already stated, min-max scaled all features into $[0, 1]$. The basic setting for the parameters was $s = 10^{-6}$, $\alpha = 10^{-3}$, and $\beta = 10^{-6}$. These were applied for all other cases except for the two-dimensional input with Border where $\alpha = 10^{-5}$ was used.

Characteristically with the alternating direction methods, key to an efficient performance of the Douglas-Rachford algorithms from the previous section lies in the selection of the timestep Δt. Typically the larger is the timestep the faster is the convergence of a splitting method, although a too large value might lead to a premature stop in a wrong solution. Here the key observation after preliminary tests was that in (3) we need to bring the additionally introduced time-dependent term and the reciprocal of Δt on the same "computational level", scale, than the penalization part of (1). Hence, a proper choice seems to have the form $\Delta t = f(m, \frac{1}{\alpha}, \frac{1}{\beta})$, which, in particular, emphasizes proper action of the first substep of the DR scheme.

When dealing with boundary value and/or initial value problems, the discretization accuracy and stability of the time-integration scheme play a central role in restricting the value of Δt. However, even in such a context, Carthel [9] observed and experimented cases where very large value of Δt gave the best rate of convergence. Hence, our suggestion for the timestep, which was elaborated and experimented heavily before the actual tests reported here, is the following: $\Delta t = 5 \cdot 10^4 \cdot \sqrt{m}$. The Douglas-Rachford algorithm was iterated as long as the relative error $\mathcal{J}(\mathbf{W}^k)/\mathcal{J}(\mathbf{W}^0)$

[1] `https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research`

Table 1: *Description of test datasets.*

Dataname	N	NT	n_0	n	Depiction
Yale32	165	-	1024	15	32x32 grayscale face images
Border	4 000	1 000	2	3	see [43, 44]
Outdoor	2 400	1 600	21	40	see [43, 44]
Landsat	4 435	2 000	36	6	32 spectral band coefficients from satellite images
USPS	7 291	2 007	256	10	Handwritten digits (10) from scanned 16x16 images
Isolet	6 238	1 559	617	26	Spoken letters (26) with acoustic features
MNIST	60 000	10 000	666	10	Handwritten digits (10) on 28x28 grayscale images

was decreasing. Maximum number of iterations was restricted to 1000.

In the tests, we incremented the number of basis functions m with the increment $m_{inc} = \lfloor N/100 \rfloor$ up to the largest size $m_{max} = \lfloor 2N/3 \rfloor$. $\mathbf{W}^0$ was generated similarly to G from the uniform distribution and, for fixed m, the random basis $\{\mathbf{h}_i\}$ was regenerated 10 times as suggested by the reviewed results in [42, 40] - as the final result reported we chose the smallest error over these ten attempts. The error was computed as *misclassifications-in-percentages, MPC*, over the test datasets given in Table **??** (except for Yale32 where we compute and illustrate the training error). As a reference method, referred as "BasicELM", we used the basic Tikhonov regularized form of the extreme learning machine in (1) with $\beta = 0$. This least-squared problem can be solved directly from (2), similarly to the second step of the DR method in Section 2.2.

3.2. Results. Illustrative dashboards of the results are given in Appendix A. For each dataset, we provide the following six figures

1) MPC-error over different values of m for BasicELM (red) and l_s (blue) regularization (figures a)
2) Number of DR iterations for l_s regularization over different values of m and ten reattempts: median number (blue) and 10% and 90% quantiles (dashed blue) (figures b)
3) Relative componentwise difference $\|\mathbf{W}_1 - \mathbf{W}_s\|_\infty / \|\mathbf{W}_s\|_F$ between the results with l_1 and l_s regularization (figures c)
4) Fröbenius norms $\|\mathbf{W}\|_F$ versus MPC-errors: BasicELM (red) and l_s (blue) regularization (figures d)
5) Weight distribution in logarithmic scale for l_s regularization with the best value of m in Table **??** (figures e)
6) Weight distribution in logarithmic scale for BasicELM with the best value of m in Table **??** (figures f)

From the numbers of iterations of the Douglas-Rachford method for solving the nonconvex problem with the l_s regularization we conclude good efficiency of the proposed algorithm. In all cases except the Border with $\alpha = 10^{-3}$, very low number of iterations was sufficient for larger problems. For the largest problem MNIST, we needed larger number of DR iterations longer, for larger relative sizes of m, compared

Table 2: *Results for the test datasets.*

	BasicELM			l_s regularization		
Dataname	m	MPC	F-norm	m	MPC	F-norm
Yale32	88	0.0	448.1	88	0.0	20.8
Border	770	7.9	4876.7	2490	7.8	2776.0
Outdoor	370	40.6	1874.9	538	36.8	148.7
Landsat	820	11.3	384.9	1405	10.7	127.3
USPS	2784	4.5	9.3	1762	4.3	6.9
Isolet	1585	6.0	775.1	2278	5.3	54.1
MNIST	18010	1.8	2.6	27010	1.8	3.1

to the other datasets. We verified with a reference solver that also the nonconvex problem was successfully solved in the experiments.

Sizes of the best networks, their accuracies, and Fröbenius-norms of the best weight matrices are summarized in Table **??** for the BasicELM and for l_s regularized problem. From these results and from the corresponding figures in Appendix A we conclude the following: for most cases (especially Yale32, Outdoor, Landsat, Isolet) the nonconvex, sparsity favoring regularization yielded to strictly and significantly smaller set of active nonzero components compared to the BasicELM. In these cases, the MPC-error was also smaller for the l_s regularization. However (see figures c), when comparing the l_s and l_1 regularization formulations, the relative difference between the weights of the corresponding networks remained always very small, of the order $10^{-3} - 10^{-4}$.

Except for USPS and MNIST (see figures d in in Appendix A), complexity of the most accurate ELM in the Fröbenius-norm was much smaller for the l_s compared to the BasicELM. The original inputs (before min-max scaling) for USPS and MNIST are composed of discrete greyscale values of input images, and as shown in Table **??** and in subfigures (e) and (f) of Figures 6 and 8, these datasets do not need large weights to combine the random kernel. Moreover, for the three datasets (USPS, Isolet, and MNIST) with the larger input dimension, the training error of the l_s regularized ELM converged to zero during training. Also except for USPS, the best network with the l_s regularization used larger number of basis functions with smaller complexity compared to BasicELM. Altogether, the sparse regularization turned out useful to obtain simpler classifiers and even more accurate classifiers in cases when there are some errors in the training data. If the classifier is just interpolating the training data to obtain small error in the test data, like in USPS and MNIST, then restricting the flexibility of the model using nonsmooth regularization does not pay off.

4. Conclusions. We proposed a novel alternating direction method for a nonsmooth and nonconvex discrete optimization problem in machine learning, for a supervised technique referred as extreme learning machine, to identify weights of a linear combination of randomly generated sigmoidal kernel. The method was based on Douglas-Rachford approach with the general shrinkage operator, which yielded to a unified algorithm for $0 \leq s \leq 1$.

Our computational experiments confirmed that a bold selection of exceptionally large value of the timestep Δt, similarly to [9], yielded to efficient convergence properties of the algorithm. Concerning the role of the nonconvex regularization to assure

the most sparse network, our results were mixed. On one hand, we obtained simpler models with good accuracy compared to the classical, basic ELM with most of the datasets. On the other hand, the relative difference between the weights obtained with the l_1 and l_s regularization remained small. Hence, our results seem to support the results given by Donoho [18], although more experiments to thoroughly compare these two formulations should be performed in the future.

REFERENCES

[1] P. Angelov and A. Sperduti, *Challenges in deep learning*, in "Proceedings of the 24th European symposium on artificial neural networks (ESANN 2016)", pp. 489–496, 2016.

[2] C. Bao, B. Dong, L. Hou, Z. Shen, X. Zhang, and X. Zhang, *Image restoration by minimizing zero norm of wavelet frame coefficients*, Inverse problems, 32:11(2016), 115004.

[3] H. H. Bauschke and W. M. Moursi, *On the order of the operators in the Douglas-Rachford algorithm*, Optimization Letters, 10:3 (2016), pp. 447–455.

[4] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM journal on imaging sciences, 2:1 (2009), pp. 183–202.

[5] J. Bect, L. Blanc-Féraud, G. Aubert, and A. Chambolle, *A l^1-unified variational framework for image restoration*, in T. Pajdla and J. Matas, editors, Computer Vision - ECCV 2004, pp. 1–13, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[6] T. Blumensath, M. Yaghoobi, and M. E. Davies, *Iterative hard thresholding and L_0 regularisation*, in "Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on", volume 3, pages III – 877–880. IEEE, 2007.

[7] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, *Occam's razor*, Information processing letters, 24:6 (1987), pp. 377–380.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends® in Machine Learning, 3:1 (2011), pp. 1–122.

[9] C. A. Carthel, *Numerical Methods For Some Exact And Approximate Controllability Problems For The Heat Equation*, PhD thesis, University of Houston, Department of Mathematics, 1995.

[10] R. Chartrand, *Nonconvex regularization for shape preservation*, in "IEEE International Conference on Image Processing (ICIP 2007), volume 1, pages I – 293–296. IEEE, 2007.

[11] R. Chartrand, *Nonconvex splitting for regularized low-rank + sparse decomposition*, IEEE Transactions on Signal Processing, 60:11 (2012), pp. 5810–5819.

[12] R. Chartrand, *Shrinkage mappings and their induced penalty functions*, in "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)", pp. 1026–1029. IEEE, 2014.

[13] R. Chartrand and B. Wohlberg, *A nonconvex ADMM algorithm for group sparsity with sparse groups*, in "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)", pp. 6009–6013. IEEE, 2013.

[14] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM review, 43:1 (2001), pp. 129–159.

[15] Y. Chen, W. Hager, F. Huang, D. Phan, X. Ye, and W. Yin, *Fast algorithms for image reconstruction with application to partially parallel MR imaging*, SIAM Journal on Imaging Sciences, 5:1 (2012), pp. 90–118.

[16] P. L. Combettes and J.-C. Pesquet, *A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery*, IEEE Journal of Selected Topics in Signal Processing, 1:4 (2007), pp. 564–574.

[17] B. Dong and Y. Zhang, *An efficient algorithm for l_0 minimization in wavelet frame based image restoration*, Journal of Scientific Computing, 54:2-3 (2013), pp. 350–368.

[18] D. L. Donoho, *For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution*, Communications on pure and applied mathematics, 59:6 (2006), pp. 797–829.

[19] J. Douglas and H. H. Rachford, *On the numerical solution of heat conduction problems in two and three space variables*, Transactions of the American mathematical Society, 82:2 (1956), pp. 421–439.

[20] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Mathematical Programming, 55:1 (1992), pp. 293–318.

[21] M.-J. Fadili and J.-L. Starck, *Monotone operator splitting for optimization problems in sparse recovery*, in "16th IEEE International Conference on Image Processing (ICIP'2009)", pp. 1461–1464. IEEE, 2009.

[22] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary Value Problems*, North-Holland, Amsterdam, 1983.

[23] A. Frank and A. Asuncion, UCI machine learning repository, 2010.

[24] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, volume 1. Springer series in statistics, New York, 2001.

[25] D. Gabay and B. Mercier, *A dual algorithm for the solution of nonlinear variational problems via finite element approximation*, Computers & Mathematics with Applications, 2:1 (1976), pp. 17 – 40.

[26] C. Gallicchio, J. D. Martin-Guerrero, A. Micheli, and E. Soria-Olivas, *Randomized machine learning approaches: Recent developments and challenges*, in "Proceedings of the 25th European Symposium on Artificial Neural Networks (ESANN 2017), pp. 77–86, 2017.

[27] R. Glowinski, T. Kärkkäinen, and K. Majava, *On the convergence of operator-splitting methods*, in Y. Kuznetsov, P. Neittaanmäki, O. Pironneau, and E. Heikkola, editors, *Numerical Methods for Scientific Computing. Variational problems and applications*, pp. 67–79, Barcelona, 2003. CIMNE.

[28] R. Glowinski and A. Marrocco, *Sur l'approximation par éléments finis d'ordre un, et la résolution par pénalisation-dualité d'une classe de problemes de Dirichlet non linéaires*, RAIRO Anal. Num ér., 9 (1975), pp. 41–76.

[29] R. Glowinski, S. J. Osher, and W. Yin, *Chapter 1 - introduction*, in R. Glowinski, S. J. Osher, and W. Yin, editors, Splitting Methods in Communication, Imaging, Science, and Engineering, pp. 1–17. Springer International Publishing Switzerland, 2016.

[30] R. Glowinski, S. J. Osher, and W. Yin, editors, *Splitting Methods in Communication, Imaging, Science, and Engineering*, Springer International Publishing Switzerland, 2016.

[31] R. Glowinski, T.-W. Pan, and X.-C. Tai, *Chapter 2 - some facts about operator-splitting and alternating direction methods*, in R. Glowinski, S. J. Osher, and W. Yin, editors, Splitting Methods in Communication, Imaging, Science, and Engineering, pp. 19–94. Springer International Publishing Switzerland, 2016.

[32] R. Glowinski and P. L. Tallec, *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*, SIAM, Philadelphia, 1989.

[33] J. Hämäläinen, S. Jauhiainen, and T. Kärkkäinen, *Comparison of internal clustering validation indices for prototype-based clustering*, Algorithms, 10:3 (2017), pp. 105–119.

[34] B. He and X. Yuan, *On the convergence rate of Douglas-Rachford operator splitting method*, Mathematical Programming, 153:2 (2015), pp. 715–722.

[35] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, *Extreme learning machine for regression and multiclass classification*, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42:2 (2012), pp. 513–529.

[36] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, *Extreme learning machine: theory and applications*, Neurocomputing, 70:1 (2006), pp. 489–501.

[37] T. Kärkkäinen, *MLP in layer-wise form with applications to weight decay*, Neural Computation, 14:6 (2002), pp. 1451–1480.

[38] T. Kärkkäinen, *On Cross-Validation for MLP Model Evaluation*, pp. 291–300, Springer Berlin Heidelberg, 2014.

[39] T. Kärkkäinen and E. Heikkola, *Robust formulations for training multilayer perceptrons*, Neural Computation, 16:4 (2004), pp. 837–862.

[40] S. Lin, X. Liu, J. Fang, and Z. Xu, *Is extreme learning machine feasible? a theoretical assessment (part ii)*, IEEE Transactions on Neural Networks and Learning Systems, 26:1 (2015), pp. 21–34.

[41] P.-L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM Journal on Numerical Analysis, 16:6 (1979), pp. 964–979.

[42] X. Liu, S. Lin, J. Fang, and Z. Xu, *Is extreme learning machine feasible? a theoretical assessment (part i)*, IEEE Transactions on Neural Networks and Learning Systems, 26:1 (2015), pp. 7–20.

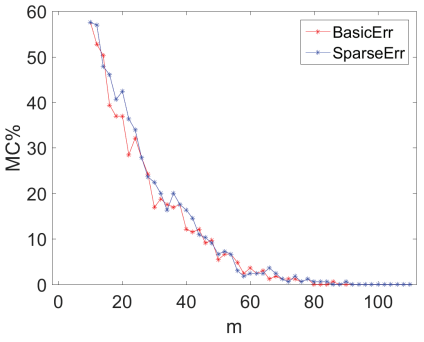
[43] V. Losing, B. Hammer, and H. Wersing, *Choosing the best algorithm for an incremental on-line learning task*, in "Proceedings of the 24th European symposium on artificial neural networks (ESANN 2016)", pp. 369–374, 2016.

[44] V. Losing, B. Hammer, and H. Wersing, *Incremental on-line learning: A review and comparison of state of the art algorithms*, Neurocomputing, 275 (2018), pp. 1261–1274.

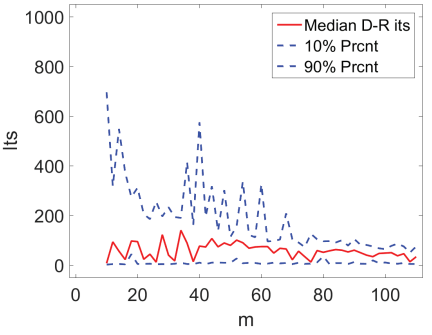
[45] M. Myllykoski, R. Glowinski, T. Kärkkäinen, and T. Rossi, *A new Augmented Lagrangian approach for L^1-mean curvature image denoising*, SIAM Journal on Imaging Sciences, 8:1

(2015), pp. 95–125.
[46] M. K. Ng, P. Weiss, and X. Yuan, *Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods*, SIAM J. Sci. Comput., 32:5 (2010), pp. 2710–2736.
[47] D. W. Peaceman and H. H. Rachford, Jr., *The numerical solution of parabolic and elliptic differential equations*, Journal of the Society for industrial and Applied Mathematics, 3:1 (1955), pp. 28–41.
[48] M. J. D. Powell, *Radial basis functions for multivariable interpolation: a review*, Algorithms for Approximation, pp. 143–167, 1987.
[49] S. Setzer, *Operator splittings, Bregman methods and frame shrinkage in image processing*, International Journal of Computer Vision, 92:3 (2011), pp. 265–280.
[50] G. Steidl and T. Teuber, *Removing multiplicative noise by Douglas-Rachford splitting methods*, Journal of Mathematical Imaging and Vision, 36:2 (2010), pp. 168–184.
[51] B. F. Svaiter, *On weak convergence of the Douglas-Rachford method*, SIAM Journal on Control and Optimization, 49:1 (2011), pp. 280–287.
[52] S. Voronin and R. Chartrand, *A new generalized thresholding algorithm for inverse problems with sparsity constraints*, in "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)", pp. 1636–1640. IEEE, 2013.
[53] Y. Wang, J. Yang, W. Yin, and Y. Zhang, *A new alternating minimization algorithm for total variation image reconstruction*, SIAM J. Imaging Sci., 1:3 (2008), pp. 248–272.
[54] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Transactions on Signal Processing, 57:7 (2009), pp. 2479–2493.
[55] J. Yang, Y. Zhang, and W. Yin, *A fast alternating direction method for TVL1-L2 signal reconstruction from partial Fourier data*, IEEE Journal of Selected Topics in Signal Processing, 4:2 (2010), pp. 288–297.
[56] X. Zhang, M. Burger, and S. Osher, *A unified primal-dual algorithm framework based on bregman iteration*, Journal of Scientific Computing, 46:1 (2011), pp. 20–46.
[57] X. Zhang and X. Zhang, *A new proximal iterative hard thresholding method with extrapolation for l_0 minimization*, Journal of Scientific Computing, pp. 1–18, 2018.
[58] Y. Zhang, B. Dong, and Z. Lu, *l_0 minimization for wavelet frame based image restoration*, Mathematics of Computation, 82:282 (2013), pp. 995–1015.
[59] W. Zhu, X.-C. Tai, and T. Chan, *Augmented Lagrangian method for a mean curvature based image denoising model*, Inverse Probl. Imag., 7:4 (2013), pp. 1409–1432.

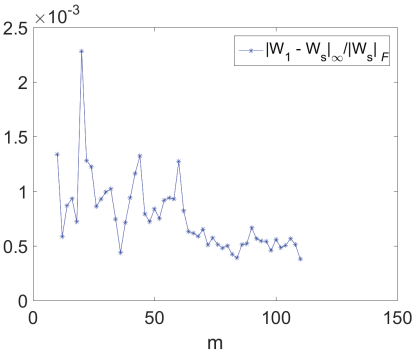
Appendix A. Result figures.



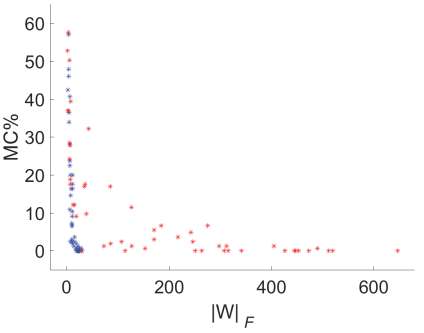
(a) Decrease of training error.



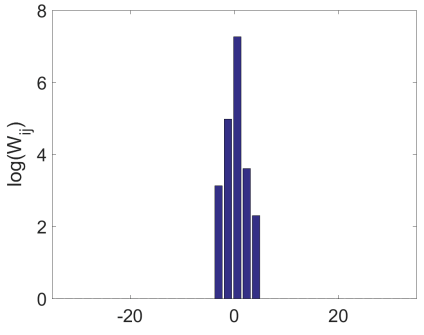
(b) Behavior of Douglas-Rachford iterations



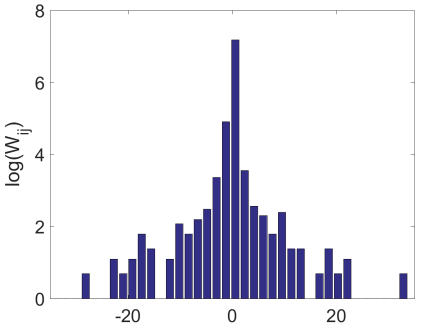
(c) Relative difference between l_1 and l_s.



(d) Scattering of $\|\mathbf{W}\|_F$ and MCP-error.

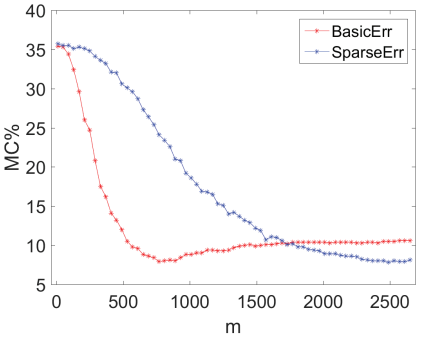


(e) Weight distribution in log-scale for l_s.

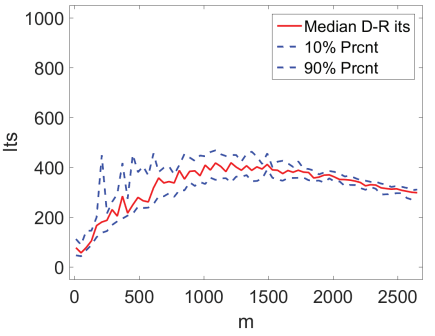


(f) Weight distribution in log-scale for BasicELM.

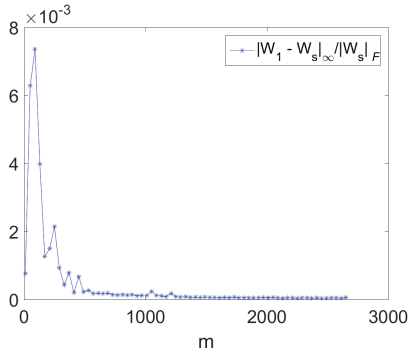
Fig. 2: *Results for Yale32.*



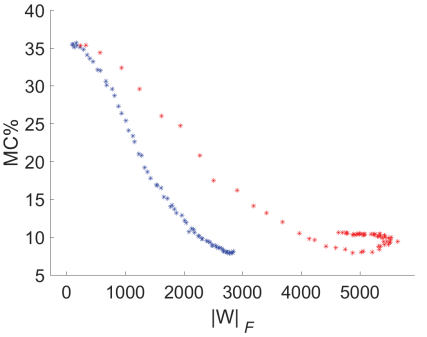
(a) Decrease of training error.



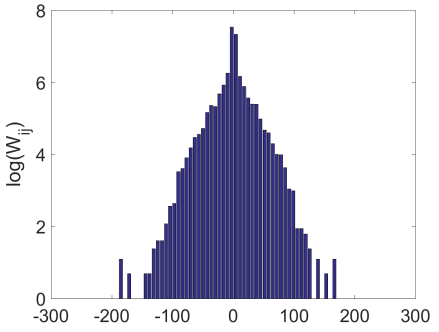
(b) Behavior of Douglas-Rachford iterations



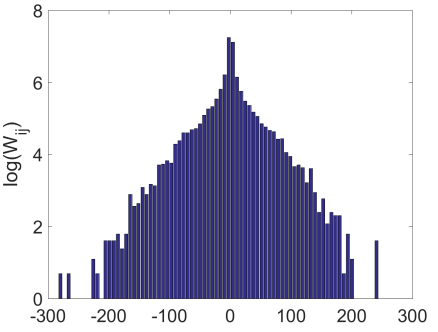
(c) Relative difference between l_1 and l_s.



(d) Scattering of $\|\mathbf{W}\|_F$ and MCP-error.

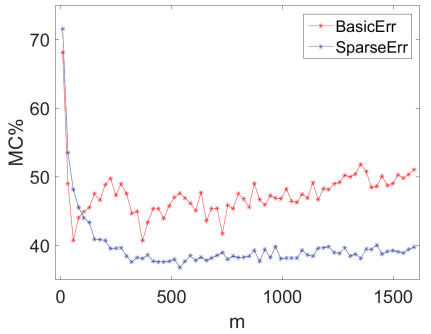


(e) Weight distribution in log-scale for l_s.

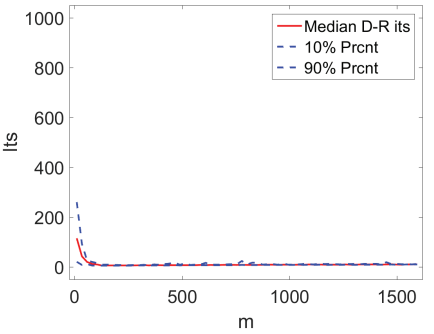


(f) Weight distribution in log-scale for BasicELM.

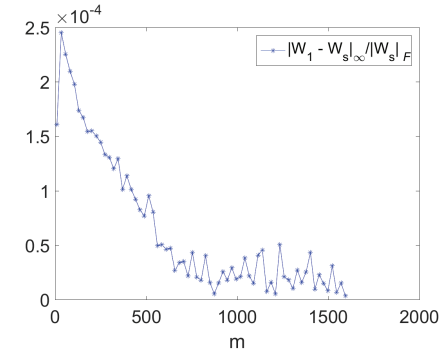
Fig. 3: *Results for Border.*



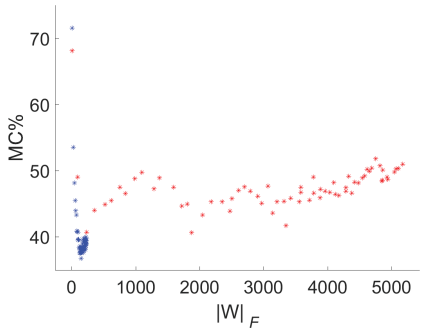
(a) Decrease of training error.



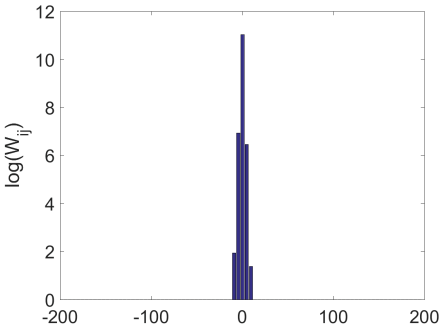
(b) Behavior of Douglas-Rachford iterations



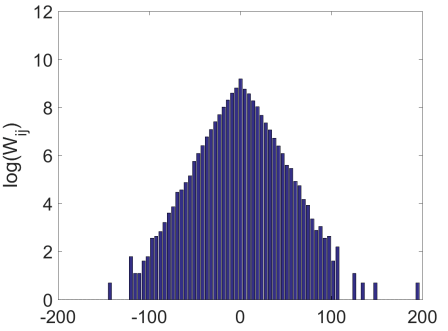
(c) Relative difference between l_1 and l_s.



(d) Scattering of $\|\mathbf{W}\|_F$ and MCP-error.

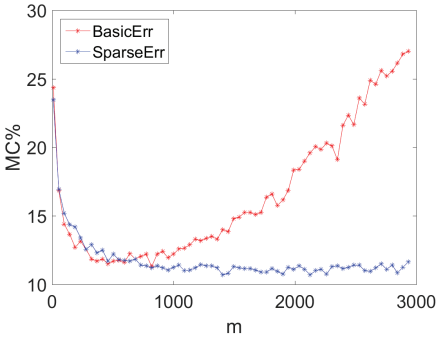


(e) Weight distribution in log-scale for l_s.

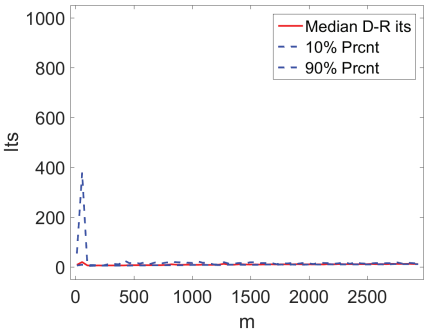


(f) Weight distribution in log-scale for BasicELM.

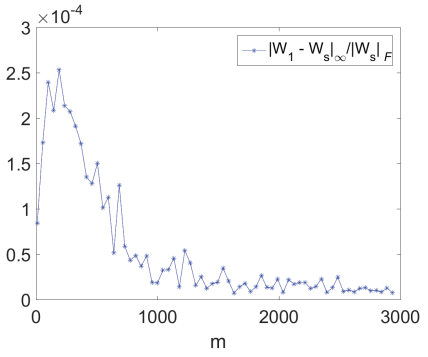
Fig. 4: *Results for Outdoor.*



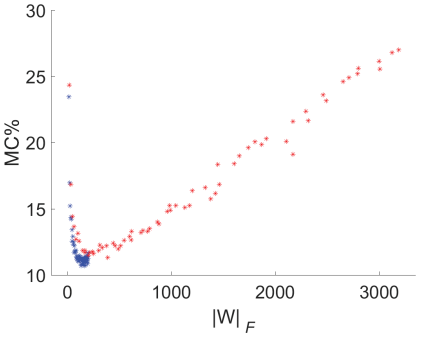
(a) Decrease of training error.



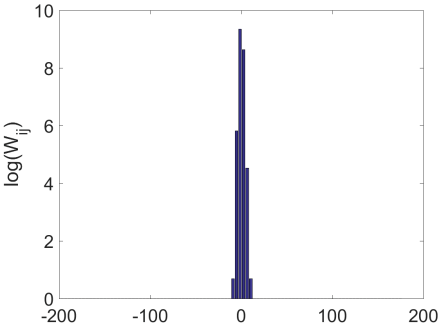
(b) Behavior of Douglas-Rachford iterations



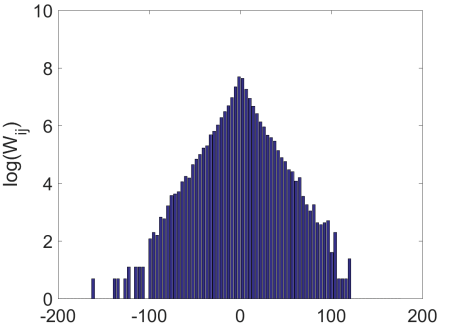
(c) Relative difference between l_1 and l_s.



(d) Scattering of $\|\mathbf{W}\|_F$ and MCP-error.

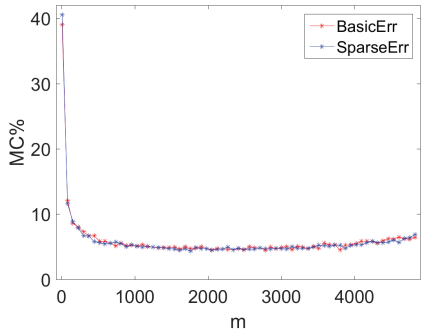


(e) Weight distribution in log-scale for l_s.

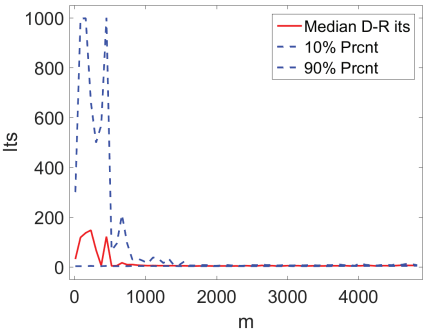


(f) Weight distribution in log-scale for BasicELM.

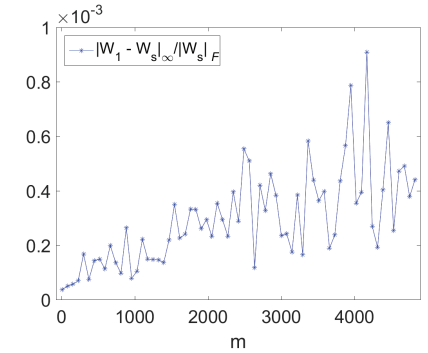
Fig. 5: *Results for Landsat.*



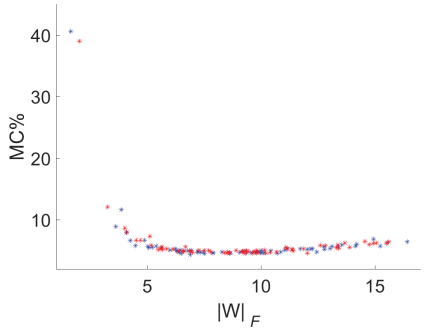
(a) Decrease of training error.



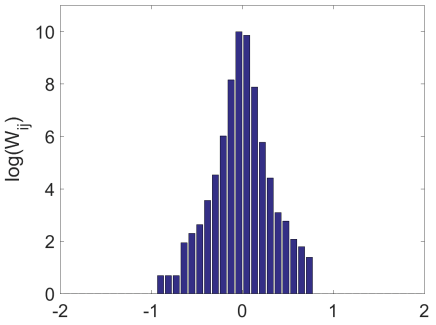
(b) Behavior of Douglas-Rachford iterations



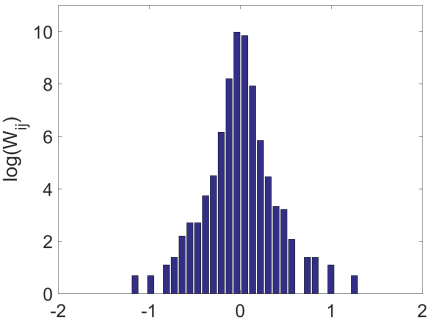
(c) Relative difference between l_1 and l_s.



(d) Scattering of $\|\mathbf{W}\|_F$ and MCP-error.

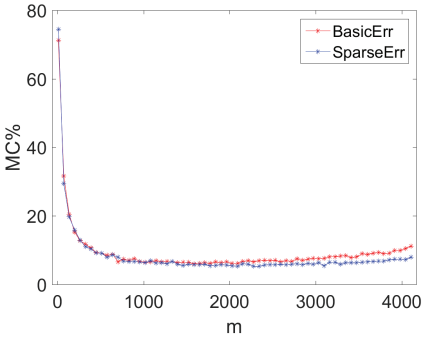


(e) Weight distribution in log-scale for l_s.

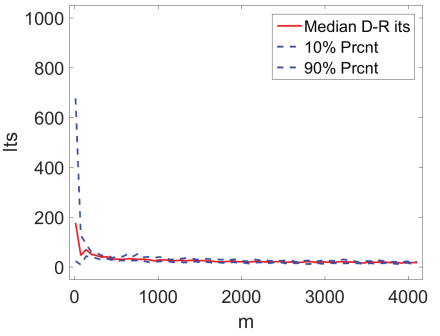


(f) Weight distribution in log-scale for BasicELM.

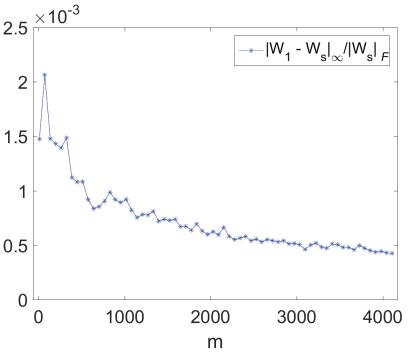
Fig. 6: *Results for USPS.*



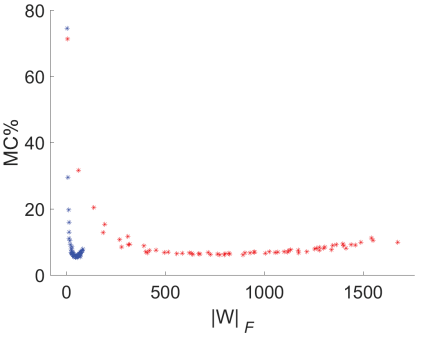
(a) Decrease of training error.



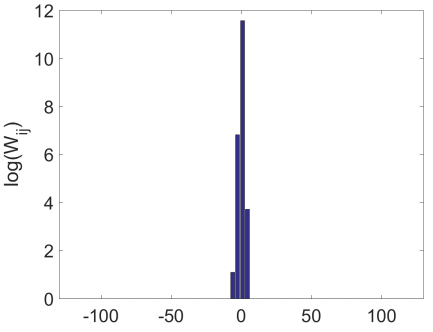
(b) Behavior of Douglas-Rachford iterations



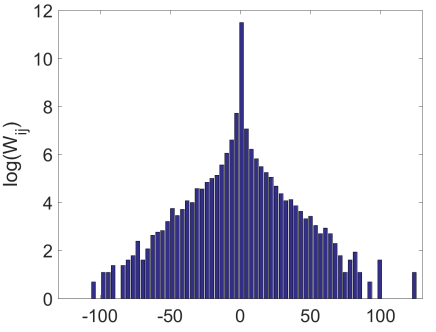
(c) Relative difference between l_1 and l_s.



(d) Scattering of $\|\mathbf{W}\|_F$ and MCP-error.

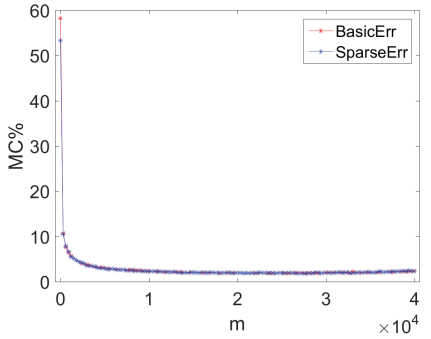


(e) Weight distribution in log-scale for l_s.

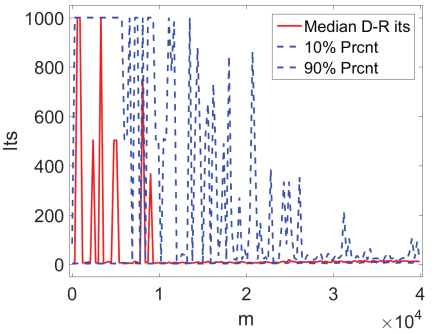


(f) Weight distribution in log-scale for BasicELM.

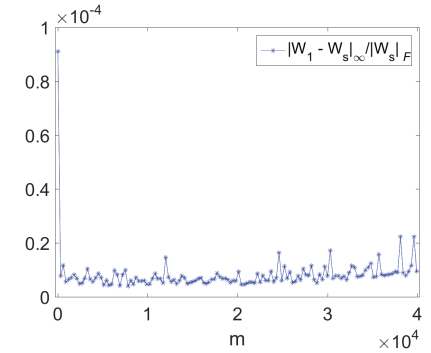
Fig. 7: *Results for Isolet.*



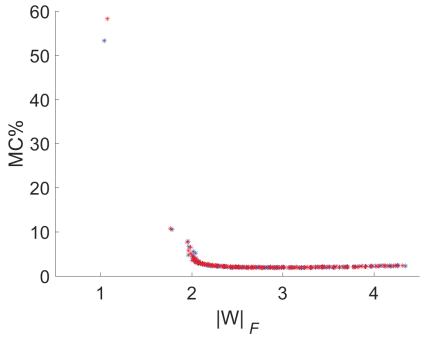
(a) Decrease of training error.



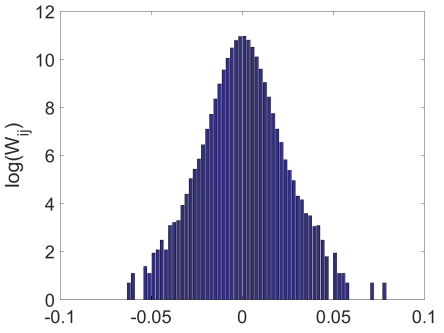
(b) Behavior of Douglas-Rachford iterations



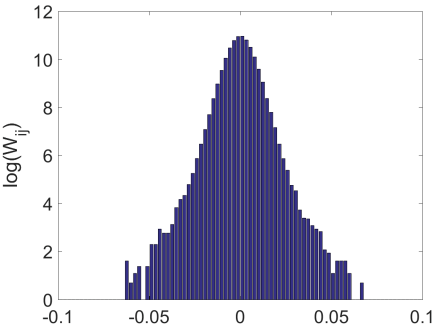
(c) Relative difference between l_1 and l_s.



(d) Scattering of $\|\mathbf{W}\|_F$ and MCP-error.



(e) Weight distribution in log-scale for l_s.



(f) Weight distribution in log-scale for BasicELM.

Fig. 8: *Results for MNIST.*