Berhu Penalty for Matrix and Tensor Estimation

Giulia Denevi[1,2], Michele Donini[1], Massimiliano Pontil[1,3]
[1]Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genova
[2]Dipartimento di Matematica, Università degli Studi di Genova, Genova
[3]Department of Computer Science, University College London, London

***Abstract*—We present a regularizer for learning low rank matrices. It is obtained by applying the Berhu penalty function [1] to the spectrum. We link the regularizer to previous ones for structured sparsity and derive its proximity operator. In numerical experiments the spectral Berhu performs favourably over the standard methods. We discuss how the regularizer can be extended to the tensor setting.**

I. Problem

We are interested in learning matrices or tensors from a set of linear measurements. Applications range from collaborative filtering, medical imaging, to natural language processing, and many more. We focus on the matrix case and later outline the extension to tensors. We prescribe a linear operator $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$, representing a set of measurements obtained from a target matrix W^* as $y = \mathcal{A}(W^*) + \eta$, where η is some disturbance noise. This framework includes various settings, depending on the choice of the operator $\mathcal{A}$, such as matrix completion and multitask learning. We attempt to recover W^* from the data $(\mathcal{A}, y)$, by solving the optimization problem

$$\min_{W \in \mathbb{R}^{d_1 \times d_2}} \left\{ \|y - \mathcal{A}(W)\|_2^2 + \gamma \Omega(W) \right\} \qquad (1)$$

where γ is a positive regularization parameter, which may be chosen by cross validation. The role of the regularizer Ω is to encourage matrices with few degrees of freedom and in this work we are interested in low rank matrices. To this end a standard convex regularizer is given by the trace (or nuclear) norm, i.e. the sum of the singular values or equivalently the ℓ_1 norm (Lasso) of the vector $\sigma(W)$ containing the singular values of W in non-increasing order, i.e. $\Omega(W) = \|\sigma(W)\|_1$.

II. Regularizer

In this work we propose a different regularizer which is related to the convex relaxation of the cardinality combined with the ℓ_2 norm of the spectrum of a matrix, and give a link to the k–support norm [2]. The regularizer is given by the convex envelope h_ϵ of the function $\mathrm{rank}(W) + \frac{\epsilon}{2}\|W\|_{\mathrm{Fr}}^2$, where ϵ is a positive parameter selected by cross validation. By von Neumann trace inequality, $h_\epsilon(W) = g_\epsilon(\sigma(W))$, where g_ϵ is the convex envelope of the function $\mathrm{card}(\cdot) + \frac{\epsilon}{2}\|\cdot\|_2^2$. The regularizer g_ϵ is equal to the Berhu function (reverse Huber) [1] and it has been motivated by [3] as a better relaxation of the Elastic Net. Fig. 1 depicts the Berhu penalty and Fig. 2 illustrates its behaviour for different choices of ϵ. The function g_ϵ can be written as a sum of univariate functions of the same kind and with some abuse of notation we write $g_\epsilon(x) = \sum_{i=1}^{\min(d_1,d_2)} g_\epsilon(x_i)$. We use a technique from [4] to express g_ϵ as an infimum of quadratics, i.e. $g_\epsilon(x_i) = \sqrt{\epsilon/2}\inf\left\{\left(\frac{x_i^2}{\theta} + \theta\right) : \theta \in \left(0, \sqrt{2/\epsilon}\right)\right\}$.

The above observation and the infimum formulation of the k–support norm [5] allow us to show that the regularization path of the Berhu function contains the regularization path of the k–support norm. Moreover, the proximity operator of γg_ϵ can be computed in a close form and is depicted in Fig. 3 for $\gamma = 1$, alongside the proximity operators of the Lasso and the Elastic Net. A natural generalization of the Berhu function is given by the convex envelope of $\mathrm{card}(\cdot) + \frac{\epsilon}{p}\|\cdot\|_p^p$, where $p > 1$ is a further parameter allowing us to better fit the spectral decay of the underlying model. In this case, the generalized Berhu is related to the (k, p)–support norm outlined in [6].

III. Numerical Experiments

We compared the spectral Berhu to the trace norm, the matrix Elastic Net and the spectral k–support norm in a low rank matrix completion problem. We reproduced the same experimental setting described in [5, Sec. 7.1]. The averaged results are shown in Table I, where the Berhu penalty, similar to the spectral k–support norm, outperforms the other methods.

Ongoing work is studying the practical value of this approach for tensor estimation. In this case W is a $d_1 \times \cdots \times d_N$ tensor and $\mathcal{A} : \mathbb{R}^{d_1 \times \cdots \times d_N} \to \mathbb{R}^m$. Following [7] we may use the sum of the Berhu regularizer g_ϵ applied to the spectrum of the matricizations, i.e. we may take $\Omega(W) = \sum_{n=1}^N \mathrm{g}_\epsilon(\sigma(W_{(n)}))$, where $W_{(n)}$ is the n^{th} matricization of the tensor W. The matrix case is recovered for $N = 2$.

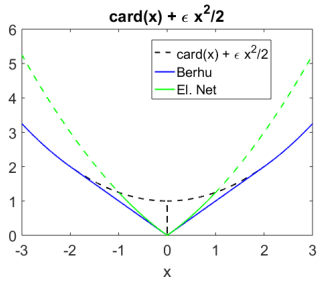

Fig. 1: Berhu and the corresponding Elastic Net penalty ($\epsilon = 0.5$).

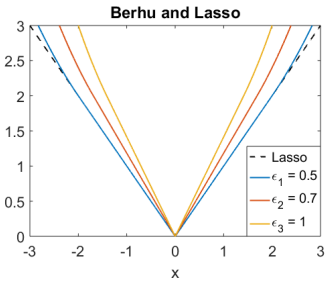

Fig. 2: Berhu for different parameters ϵ ($\epsilon_1 < \epsilon_2 < \epsilon_3$).

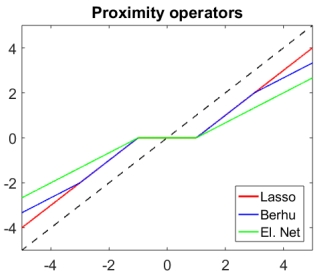

Fig. 3: Proximity operators: Lasso is not able to shrink large values to preserve the data correlation, encouraged by the Elastic Net; Berhu penalty has the advantage of shifting medium and shrinking large values in a separate way ($\epsilon = 0.5$).

TABLE I: Matrix completion on simulated data (100 trials): $W = A^T B + E$, where $A, B \in \mathbb{R}^{100\times r}$ (with $r = 5, 10$), $E \in \mathbb{R}^{100\times 100}$. The entries were taken i.i.d. Gaussian: 10% of them as validation set and varying the percentage for the training set.

Rank	Training	Regularizer	Test error
5	10 %	Trace	0.7887 ± 0.03
		Elastic Net	0.7897 ± 0.03
		k–supp.	0.7844 ± 0.03
		Berhu	$\mathbf{0.7837} \pm 0.03$
5	15 %	Trace	0.5645 ± 0.04
		Elastic Net	0.5630 ± 0.04
		k–supp.	0.5639 ± 0.03
		Berhu	$\mathbf{0.5617} \pm 0.04$
5	20 %	Trace	0.4081 ± 0.03
		Elastic Net	0.4069 ± 0.03
		k–supp.	0.4059 ± 0.03
		Berhu	$\mathbf{0.4047} \pm 0.03$
10	20 %	Trace	0.6253 ± 0.03
		Elastic Net	0.6243 ± 0.03
		k–supp.	0.6239 ± 0.03
		Berhu	$\mathbf{0.6225} \pm 0.03$
10	30 %	Trace	0.3648 ± 0.02
		Elastic Net	0.3635 ± 0.02
		k–supp.	0.3615 ± 0.02
		Berhu	$\mathbf{0.3601} \pm 0.02$
10	40 %	Trace	0.2240 ± 0.02
		Elastic Net	0.2235 ± 0.02
		k–supp.	0.2209 ± 0.02
		Berhu	$\mathbf{0.2198} \pm 0.02$
10	50 %	Trace	0.1533 ± 0.01
		Elastic Net	0.1533 ± 0.01
		k–supp.	$\mathbf{0.1473} \pm 0.01$
		Berhu	0.1475 ± 0.01

REFERENCES

[1] A.B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

[2] A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the k-support norm. *Advances in Neural Information Processing Systems 25*, pp. 1466–1474, 2012.

[3] V. Jojic, S. Saria, and D. Koller. Convex envelopes of complexity controlling penalties: the case against premature envelopment. In *Proc. 14th Int. Conf. on Artificial Intelligence and Statistics*, pp. 399–406, 2011.

[4] C. A. Micchelli, J. M. Morales, and M. Pontil. Regularizers for structured sparsity. *Advances in Comp. Mathematics*, 38:455–489, 2013.

[5] A. McDonald, M. Pontil, and D. Stamos. New perspectives on k-support and cluster norms. *Journal of Machine Learning Research*, 17(155):1–38, 2016.

[6] A. McDonald, M. Pontil, and D. Stamos. Fitting spectral decay with the k-support norm. In *Proc. 19th Int. Conf. on Artificial Intelligence and Statistics*, pp. 1061–1069, 2016.

[7] B. Romera-Paredes and M. Pontil. A new convex relaxation for tensor completion. In *Advances in Neural Information Processing Systems*, pp. 2967–2975, 2013.