Comment on the subdifferential formula given in the article "Algorithmic Analysis and Statistical Estimation of SLOPE via Approximate Message Passing"

Patrick Tardivel, Institut de Mathématiques de Bourgogne, UMR 5584 CNRS, Université de Bourgogne, 9 avenue Alain Savary, Dijon, 21078, France

Abstract

Subdifferential formula for the sorted ℓ_1 norm attracted lot of attention recently to derive screening procedures for SLOPE [5, 7], clustering and sparsity properties for SLOPE [1, 10] and explicit expressions for the proximal operator of the sorted ℓ_1 norm [4, 11]. In this note, we discuss the formula for the subdifferential of the sorted ℓ_1 norm given in the article of Bu et al., [3]. We believe that expressing the sorted ℓ_1 norm as the maximum of linear functions, as suggested by the authors, is an appropriate approach to derive its subdifferential. However the formula provided in this article contains many flaws and we hope that authors would take into account this note in order to rewrite and prove their formula.

We remind that the sorted ℓ_1 norm is defined as follows

$$J_\lambda(x) = \sum_{i=1}^{p} \lambda_i |x|_{(i)},$$

where $\lambda = (\lambda_1, \dots, \lambda_p) \in \mathbb{R}^p$ with $\lambda_1 > 0$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ is a sequence of penalty parameters and $|x|_{(1)} \geq \cdots \geq |x|_{(p)}$ are non-increasing components of x in absolute value. Fact V3 in the article of Bu et al., [3] provides a formula for the subdifferential of J_λ reported hereafter[1]:

$$\partial J_\lambda(s) = \left\{ v \in \mathbb{R}^p : \text{ for each equivalent class } I, \begin{cases} \text{if } s_I \neq 0 \Longrightarrow v_I \in \mathcal{P}([\hat{\Pi}_s^{-1}(\lambda)]_I)\text{sgn}(s_I) \\ \text{if } s_I = 0 \Longrightarrow v_I \in \mathcal{P}_0([\hat{\Pi}_s^{-1}(\lambda)]_I) \end{cases} \right\}. \quad (1)$$

In the above, $\mathcal{P}$, $\mathcal{P}_0$ are polytope-related mappings:

$$\mathcal{P}(u) = \{y : y = Au \text{ for some doubly stochastic matrix } A\}$$

$$\mathcal{P}_0(u) = \{y : y = Au \text{ for some doubly sub-stochastic matrix } A\}$$

Hereafter, we are going to discuss the notation $\hat{\Pi}_s^{-1}$ given in (1), the polytope mappings as well as sketch of proof for this formula. First, before commenting (1), we remind that for a given norm $\|\cdot\|$

[1] This formula also appears in supplementary material of the article [2]; derivation for the subdifferential of the sorted ℓ_1 norm are very similar.

whose dual norm is $\|\cdot\|^*$, the subdifferential at s satisfies (see *e.g.* [6])

$$\partial\|\cdot\|(s) = \{v \in \mathbb{R}^p : \|v\|^* \leq 1 \text{ and } v^T s = \|s\|\}.$$

In particular

- The subdifferential at 0 is the unit ball of the dual norm.
- If v is an arbitrary subgradient of $\|\cdot\|$ at $s \in \mathbb{R}^p$, namely $v \in \partial\|\cdot\|(s)$, then $\|v\|^* \leq 1$.

Note that the dual sorted ℓ_1 norm has the following explicit expression (see *e.g.* [8])

$$J_\lambda^*(x) = \max\left\{\frac{|x|_{(1)}}{\lambda_1}, \ldots, \frac{|x|_{(1)} + |x|_{(2)}}{\lambda_1 + \lambda_2}, \ldots, \frac{|x|_{(1)} + \cdots + |x|_{(p)}}{\lambda_1 + \cdots + \lambda_p}\right\}.$$

Mappings $\hat{\Pi}_x$ and $\hat{\Pi}_x^{-1}$

The mapping $\hat{\Pi}_x : \mathbb{R}^p \to \{\text{maximal atoms}\}$ and its pseudo-inverse $\hat{\Pi}_x^{-1}$ are not properly defined; instead authors only provide the following example for these notations. If $x = (5, 2, -3, -5)$ then $\hat{\Pi}_x(x) = (\{5, -5\}, \{5, -5\}, -3, 2)$ and $\hat{\Pi}_x^{-1}(\lambda) = (\{\lambda_1, \lambda_2\}, \lambda_4, \lambda_3, \{\lambda_1, \lambda_2\})$. As pointed out by the authors, there exists $\hat{\lambda} \in \hat{\Pi}_x^{-1}(\lambda)$ such that $J_\lambda(x) = \langle\hat{\lambda}, |x|\rangle$[2]. We agree with this fact, indeed

$$J_\lambda(x) = \lambda_1|x_1| + \lambda_4|x_2| + \lambda_3|x_3| + \lambda_2|x_4| = \lambda_2|x_1| + \lambda_4|x_2| + \lambda_3|x_3| + \lambda_1|x_4|.$$

Therefore, $J_\lambda(x) = \langle\hat{\lambda}, |x|\rangle$ as soon as $\hat{\lambda} \in \{(\lambda_1, \lambda_4, \lambda_3, \lambda_2), (\lambda_2, \lambda_4, \lambda_3, \lambda_1)\}$. This fact suggests that $\hat{\Pi}_x^{-1}(\lambda)$ is a Cartesian product, namely

$$\hat{\Pi}_x^{-1}(\lambda) = \{(\lambda_1, \lambda_4, \lambda_3, \lambda_1), (\lambda_1, \lambda_4, \lambda_3, \lambda_2), (\lambda_2, \lambda_4, \lambda_3, \lambda_1), (\lambda_2, \lambda_4, \lambda_3, \lambda_2)\}.$$

However, if $\hat{\Pi}_x^{-1}(\lambda)$ is Cartesian product then the formula given in (1) is wrong since $(\lambda_1, \lambda_4, \lambda_3, \lambda_1) \notin \partial J_\lambda(x)$ (indeed $J_\lambda^*((\lambda_1, \lambda_4, \lambda_3, \lambda_1)) > 1$). Overall we believe that the notations $\hat{\Pi}_x$ and $\hat{\Pi}_x^{-1}$ are very complicated and not useful to derive a formula for the subdifferential of J_λ. Instead, one may probably replace $\hat{\Pi}_s^{-1}(\lambda)$, in (1), by $\hat{\lambda} = (\lambda_{\pi^{-1}(1)}, \ldots, \lambda_{\pi^{-1}(p)})$ where π^{-1} is the inverse of π an arbitrary permutation in $\{1, \ldots, p\}$ for which $(|x_{\pi(1)}|, \ldots, |x_{\pi(p)}|) = (|x|_{(1)}, \ldots, |x|_{(p)})$.

Polytope mappings

We agree with the authors, when components of s are all equal (and positive) then $\partial J_\lambda(s)$ is the permutoèdre: $\text{conv}\{(\lambda_{\pi(1)}, \ldots, \lambda_{\pi(p)}) : \text{ where } \pi \text{ is a permutation on } \{1, \ldots, p\}\}$. The permutoèdre is closely related to the Birkhoff polytope (the set of doubly stochastic matrices); indeed,

$$\mathcal{P}(\lambda) = \text{conv}\{(\lambda_{\pi(1)}, \ldots, \lambda_{\pi(p)}) : \text{ where } \pi \text{ is a permutation on } \{1, \ldots, p\}\}.$$

[2]This statement is not formulated this way; in the article there is a confusion between x and b.

On the other hand, the link between the unit ball of the dual norm J_λ^* (the permutoèdre signé) and the set of the sub-stochastic matrices is unclear. We conjecture the following statement

$$\mathcal{P}_0(\lambda) = \{x \in [0, +\infty)^p : J_\lambda^*(x) \leq 1\}.$$

However, if this conjecture is wrong it would be impossible to recover the permutoèdre signé: $\partial J_\lambda(0)$, based on $\mathcal{P}_0$.

Sketch of proof for the formula

Authors claim that a rigorous proof of Fact V3 is given in [9, Exercise 8.31]. Actually, this exercise only provides a well known formula for the subdifferential of the maximum of a finite family of convex smooth functions (this fact is also proved in the book of Hiriart-Urruty and Lemaréchal [6, pp. 182-183,187-188]). So, currently fact V3 is not yet proven. Concerning the sketch of proof, authors "rewrite $J_\lambda(s)$ as a finite max function

$$J_\lambda(s) = \max\{\lambda^T f_1(s), \dots, \lambda^T f_m(s)\}, \tag{2}$$

where $\{f_i(s)\}_{1 \leq i \leq m}$ is the collection of all possible permutations for the entries of $|s|$". To our understanding expression (2) is unclear, indeed:

1. If $f_i(s) = (|s_{\pi(1)}|, \dots, |s_{\pi(p)}|)$ for some permutation π on $\{1, \dots, p\}$ then the formula (2) is true. However, $\lambda^T f_i(s)$ is a weighted ℓ_1 norm thus not a smooth function, the gradient[3] $\nabla_s \lambda^T f_i(s)$ depends on $\operatorname{sgn}(s) \in \{-1, 1\}^p$ and the notation $f_i^{-1}(\lambda)$ does not make sense.
2. If $f_i(s) = (s_{\pi(1)}, \dots, s_{\pi(p)})$ then the gradient satisfies $\nabla_s \lambda^T f_i(s) = f_i^{-1}(\lambda)$. However, the formula (2) is wrong as soon as some components of s are negative.

Authors claims that when components of s are all equal then

$$\partial J_\lambda(s) = \operatorname{conv}\{(\lambda_{\pi(1)}, \dots, \lambda_{\pi(p)}) : \text{ where } \pi \text{ is a permutation on } \{1, \dots, p\}\}.$$

This equality is true when components of s are all equal and positive (see *e.g* [4, 10]). However, this formula is no longer true when components of s are all equal and negative[4]. Finally, we believe that the formula $\partial J_\lambda(s) = \operatorname{conv}\{f_i^{-1}(\lambda) : i \in A(s)\}$ is misleading. Indeed, if $f_i^{-1}(\lambda)$ is just a permutation of components of λ then we would have the following inclusion

$$\partial J_\lambda(s) = \operatorname{conv}\{f_i^{-1}(\lambda) : i \in A(s)\} \subset \operatorname{conv}\{(\lambda_{\pi(1)}, \dots, \lambda_{\pi(p)}) : \text{ where } \pi \text{ is a permutation on } \{1, \dots, p\}\}.$$

However this inclusion is wrong as soon as some components of s are negative. To conclude, instead to use implicitly the permutation group, we suggest to authors to use the signed permutation group G (see *e.g.* [8] or [10, Definition 22]) and to rewrite the sorted ℓ_1 has a maximum of linear functions:

$$J_\lambda(s) = \max\{g(\lambda)^T s : g \in G\}.$$

[3] The weighted ℓ_1 norm is differentiable at s since, by assumption, s does not have any null component.
[4] This formula is also wrong when $s = 0$ but this particular case is discarded by the authors.

This formula might be very useful to derive the subdifferental of J_λ.

References

[1] Małgorzata Bogdan, Xavier Dupuis, Piotr Graczyk, Bartosz Kołodziejek, Tomasz Skalski, Patrick Tardivel, and Maciej Wilczyński. Pattern recovery by slope. *arXiv preprint arXiv:2203.12086*, 2022.

[2] Zhiqi Bu, Jason Klusowski, Cynthia Rush, and Weijie Su. Algorithmic analysis and statistical estimation of slope via approximate message passing. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] Zhiqi Bu, Jason M Klusowski, Cynthia Rush, and Weijie J Su. Algorithmic analysis and statistical estimation of slope via approximate message passing. *IEEE Transactions on Information Theory*, 67(1):506–537, 2020.

[4] Xavier Dupuis and Patrick JC Tardivel. Proximal operator for the sorted ℓ_1 norm: Application to testing procedures based on slope. *Journal of Statistical Planning and Inference*, 221:1–8, 2022.

[5] Clément Elvira and Cédric Herzet. Safe rules for the identification of zeros in the solutions of the slope problem. *SIAM Journal on Mathematics of Data Science*, 5(1):147–173, 2023.

[6] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

[7] Johan Larsson, Malgorzata Bogdan, and Jonas Wallin. The strong screening rule for slope. *Advances in neural information processing systems*, 33:14592–14603, 2020.

[8] Renato Negrinho and Andre Martins. Orbit regularization. *Advances in neural information processing systems*, 27, 2014.

[9] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[10] Ulrike Schneider and Patrick Tardivel. The geometry of uniqueness, sparsity and clustering in penalized estimation. *Journal of Machine Learning Research*, 23(331):1–36, 2022.

[11] Patrick JC Tardivel, Rémi Servien, and Didier Concordet. Simple expressions of the lasso and slope estimators in low-dimension. *Statistics*, 54(2):340–352, 2020.