A response to "Fast OSCAR and OWL Regression via Safe Screening Rules" by Bao et al.

Clément Elvira, Cédric Herzet

A response to "Fast OSCAR and OWL Regression via Safe Screening Rules" by Bao *et al.*

Clément Elvira and Cédric Herzet

I. Introduction

In this note, we discuss a recent contribution [2] of Bao *et al.* which addressed the so-called *"Ordered Weighted L-One Linear Regression"* (OWL) problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} (\mathbf{x}) \triangleq \tfrac{1}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 + r_{\text{OWL}}(\mathbf{x}) \qquad \text{(1 - OWL)}$$

where

$$r_{\text{OWL}}(\mathbf{x}) \triangleq \sum_{k=1}^{n} \gamma_k |\mathbf{x}|_{[k]} \qquad (2)$$

with

$$\gamma_1 > 0, \qquad \gamma_1 \geq \cdots \geq \gamma_n \geq 0, \qquad \text{(H-3)}$$

and $|\mathbf{x}|_{[k]}$ refers to the kth largest element of $\mathbf{x}$ in absolute value.

In their paper, the authors extend the so-called "safe screening" methodology introduced by El Ghaoui [5] to the OWL problem. The concept of safe screening has already been well-studied in the literature for "sparsity-promoting" convex problems (see [6] and references therein): it consists in devising simple "tests" that allow to detect zero entries in the solution of an optimization problem. The terminology "safe" refers to the fact that all the elements passing the test are *guaranteed* to correspond to zeros of the solution.

Nevertheless, we emphasize in this note that, contrarily to what is claimed by the authors, the test proposed in [2] does not satisfy the "safe" property. In particular, we provide two counterexamples in which the screening tests proposed in [2] leads to the screening of some *nonzero* coefficients. These counterexamples question the correctness of their result and suggest the presence of technical flaws in their derivations.

The note is organized as follows. In Section II, we introduce the notational conventions used throughout the document. Section III presents the rationale of the safe screening proposed by Bao *et al.* in [2] and recall their main result. Section IV discusses the correctness of this result by leveraging two counterexamples. All technical proofs are postponed to Appendices A and B.

C. Elvira is with SCEE/IETR UMR CNRS 6164, CentraleSupelec, Cesson Sévigné, France. clement.elvira@centralesupelec.fr, https://c-elvira.github.io/.
C. Herzet is with Inria centre Rennes - Bretagne Atlantique, Rennes, France. cedric.herzet@inria.fr, http://people.rennes.inria.fr/Cedric.Herzet/

II. Notational conventions

We will use the following notations throughout the paper. Vectors are denoted by lowercase bold letters (*e.g.*, $\mathbf{x}$) and matrices by uppercase bold letters (*e.g.*, $\mathbf{A}$). The "all-zero" (respectively "all-ones") vector of dimension n is written $\mathbf{0}_n$ (resp. $\mathbf{1}_n$). Similarly, $\mathbf{1}_{m\times n}$ denotes the "all-one" matrix of $\mathbb{R}^{m\times n}$. We use symbol $^{\mathrm{T}}$ to denote the transpose of a vector or a matrix. $\mathbf{x}_j$ refers to the jth component of $\mathbf{x}$. When referring to the sorted entries of a vector, we use bracket subscripts; more precisely, the notation $\mathbf{x}_{[k]}$ refers to the kth largest value of $\mathbf{x}$. For matrices, we use $\mathbf{a}_j$ to denote the jth column of $\mathbf{A}$. We use the notation $|\mathbf{x}|$ to denote the vector made up of the absolute value of the components of $\mathbf{x}$. The sign function is defined for all scalars x as $\operatorname{sign}(x) = x/|x|$ with the convention $\operatorname{sign}(x) = 0$.
Calligraphic letters are used to denote sets (*e.g.*, $\mathcal{J}$). If $a < b$ are two integers, $[\![a, b]\!]$ is used as a shorthand notation for the set $\{a, a+1, \ldots, b\}$.

For all $\mathbf{x} \in \mathbb{R}^n$ and $j \in [\![1, n]\!]$, $\mathtt{r}(\mathbf{x}, j)$ is such that $\mathbf{x}_{(j)} = \mathbf{x}_{[\mathtt{r}(\mathbf{x},j)]}$, that is it refers to the position of the jth entry of $\mathbf{x}$ when $\mathbf{x}$ is sorted. For instance, if $\mathbf{x} = [2, 3, 1, 0]$ $\mathtt{r}(\mathbf{x}, 1) = 2$, $\mathtt{r}(\mathbf{x}, 2) = 1$, $\mathtt{r}(\mathbf{x}, 3) = 3$ and $\mathtt{r}(\mathbf{x}, 4) = 4$. If some of the coordinates of $\mathbf{x}$ are equal, we break ties arbitrarily. In the sequel, we will refer to $\mathtt{r}$ as the "rank" function.

III. Rationale of Safe screening rule for OWL

In this section, we summarize the main ingredients grounding the safe screening rules for OWL proposed in [2]. To distinguish between our derivations and the results of [2], all equation numbers referring to a result of the latter paper are prefixed by "B-" (*e.g.*, (B-7)).

The screening procedure proposed in [2] leverages the solution of the Fenchel dual problem of OWL. In particular, Bao *et al.* claim that [2, Eq. (6d)]

$$\mathbf{u}^\star = \arg\max_{\mathbf{u}\in\mathcal{U}} \; D(\mathbf{u}) \triangleq -\tfrac{1}{2}\|\mathbf{u}\|_2^2 - \mathbf{u}^{\mathrm{T}}\mathbf{y}, \qquad \text{(B-4)}$$

is the (Fenchel) dual problem to (1 - OWL) where

$$\mathcal{U} = \left\{\mathbf{u}\in\mathbb{R}^m \colon \forall j \in [\![1, n]\!],\ \left|\mathbf{a}_j^{\mathrm{T}}\mathbf{u}\right| \leq \gamma_{\mathtt{r}(\mathbf{x}^\star, j)}\right\} \qquad \text{(B-5)}$$

is the dual feasible set and $\mathbf{x}^\star \in \mathbb{R}^n$ is a minimizer of (1 - OWL). At this stage, let us note that $\mathbf{u}^\star$ exists and is unique (*i.e.*, the equality in (B-4) is well-defined) because D is a continuous strongly-concave function and $\mathcal{U}$ a closed set.

Besides, $\mathbf{u}^\star$ is connected to any minimizer $\mathbf{x}^\star$ of (1 - OWL) through the identity

$$\mathbf{u}^\star = \mathbf{y} - \mathbf{A}\mathbf{x}^\star \tag{6}$$

as a consequence of standard primal-dual optimality conditions [3, Th. 19.1].

We now describe the safe screening test for OWL proposed in [2]. One easily verifies that (1 - OWL) admits at least one minimizer so that the screening problem is well-posed. Consider therefore a minimizer $\mathbf{x}^\star$ of (1 - OWL). Bao *et al.* claim that[1] [2, Eq. (11)]

$$\left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star\right| < \gamma_{r(\mathbf{x}^\star,\ell)} \quad \Longrightarrow \quad \mathbf{x}_\ell^\star = 0. \tag{B-7}$$

In other words, (B-7) describes the following sufficient condition for *safe* screening: if the inner product between the ℓth column of $\mathbf{A}$ and $\mathbf{u}^\star$ – the maximizer of the dual problem (B-4) – is lower than some threshold, the ℓth entry of any minimizer of OWL can be set to zero. We have nevertheless identified some technical flows in the derivation of (B-7) which prevent test (B-7) from being safe. This will be discussed in Section IV.

We may notice that (B-7) requires the knowledge of both $\mathbf{u}^\star$ and the ordering of a solution of OWL. On the one hand, computing $\mathbf{u}^\star$ is usually as difficult as solving (1 - OWL). On the other hand, the ordering of $\mathbf{x}^\star$ is obviously not known beforehand. As a consequence, (B-7) is of poor practical interest and Bao *et al.* devised a relaxed version of their test.

The relaxed test leverages the following two ingredients. First, since $\gamma_{r(\mathbf{x}^\star,\ell)} \geq \gamma_n$ for all index $\ell \in [\![1, n]\!]$ by (H-3), one immediately deduces that

$$\left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star\right| < \gamma_n \quad \Longrightarrow \quad \mathbf{x}_\ell^\star = 0. \tag{B-8}$$

Second, the knowledge of $\mathbf{u}^\star$ can be circumvented if one has access to a couple of primal / dual feasible vectors. In particular if $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{u} \in \mathcal{U}$ are known, the test simplifies as

$$\left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}\right| + \|\mathbf{a}_\ell\|_2\sqrt{2(P(\mathbf{x}) - D(\mathbf{u}))} < \gamma_n \implies \mathbf{x}_\ell^\star = 0, \tag{B-9}$$

see [2, Eq. (19)]. Bao *et al.* then proposed to interleave the relaxed screening test (B-9) with the iterations of several iterative algorithms to assess its effectiveness.

IV. Discussion

In this section, we discuss the arguments developped in [2] to show the safeness of test (B-8). In particular, Bao *et al.* ground their reasoning on the following optimality conditions

$$\mathbf{a}_j^{\mathrm{T}}\mathbf{u}^\star + \operatorname{sign}\left(\mathbf{x}_j^\star\right)\gamma_{r(\mathbf{x}^\star,j)} = 0 \text{ if } \mathbf{x}_j^\star \neq 0 \tag{B-10a}$$

$$\left|\mathbf{a}_j^{\mathrm{T}}\mathbf{u}^\star\right| \leq \gamma_{r(\mathbf{x}^\star,j)} \text{ otherwise,} \tag{B-10b}$$

and used the series of implications

$$\text{(B-10b)} \implies \text{(B-7)} \implies \text{(B-8)} \tag{11}$$

[1]Note that the authors did not mention how to handle the design of the rank function in presence of multiple maximizers.

to prove the safeness of (B-8). We show below that neither (B-10b) nor (B-7) characterize the minimizers of generic OWL problems, thus invalidating their proof of safeness.

We first provide a counterexample where a minimizer of OWL violates (B-7):

Counterexample 1. *Consider the OWL problem with dimensions $m, n \geq 2$, parameters $\mathbf{y} = \mathbf{1}_m$, $\mathbf{A} = \mathbf{1}_{m\times n}$ and weighting coefficients $\{\gamma_k\}_{k=1}^n$ such that*

$$1 \geq \gamma_1 > \gamma_2 > \cdots > \gamma_n > 0. \tag{12}$$

In this setup, we show in Appendix A that the unique solution $\mathbf{x}^\star$ of the OWL problem writes $\mathbf{x}^\star = x^\star\mathbf{1}_n$ where

$$x^\star = \frac{mn - \sum_{k=1}^n \gamma_k}{mn^2} > 0. \tag{13}$$

As a consequence of (6), we have

$$\mathbf{A}^{\mathrm{T}}\mathbf{u}^\star = \frac{\sum_{k=1}^n \gamma_k}{n}\mathbf{1}_n. \tag{14}$$

Finally, let $\ell \in [\![1, n]\!]$ be an index such that $r(\mathbf{x}^\star, \ell) = 1$. Using (12) we obtain

$$\left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star\right| < \gamma_1. \tag{15}$$

Hence $\mathbf{x}_\ell^\star = 0$ by (B-7) which leads to a contradiction.

Counterexample 1 describes a scenario where both the minimizer $\mathbf{x}^\star$ of OWL and the maximizer $\mathbf{u}^\star$ of the dual problem can be written in closed-form so that the safeness of screening test (B-7) can be easily assessed. We note that, although all entries of $\mathbf{x}^\star$ are positive, (at least) one entry of $\mathbf{x}^\star$ is screened by test (B-7). This shows that implication (B-7) is violated in this particular setup.

We next provide a counter-example showing that (B-10a) and (B-10b) do not correctly describe the optimality condition of (1 - OWL):

Counterexample 2. *Consider the OWL problem with dimensions $m = n \geq 2$, parameters $\mathbf{y} = \mathbf{1}_n$, $\mathbf{A} = \mathbf{I}_n$ the identity matrix of $\mathbb{R}^{n\times n}$ and weighting coefficients $\{\gamma_k\}_{k=1}^n$ satisfying*

$$\gamma_1 = 1 \text{ and } \gamma_k = 0 \ \forall k \in [\![2, n]\!]. \tag{16}$$

Then, the OWL problem (1 - OWL) reduces to

$$\min_{\mathbf{x}\in\mathbb{R}^n} \tfrac{1}{2}\|\mathbf{1}_n - \mathbf{x}\|_2^2 + \|\mathbf{x}\|_\infty. \tag{17}$$

We show in Appendix B that the solution of (17) is unique and writes

$$\mathbf{x}^\star = \frac{n-1}{n}\mathbf{1}_n. \tag{18}$$

One also has (as a consequence of (6))

$$\mathbf{u}^\star = \frac{1}{n}\mathbf{1}_n. \tag{19}$$

Hence, all entries of $\mathbf{x}^\star$ are positive. Yet, for all $j \in [\![1, n]\!]$,

$$\mathbf{a}_j^{\mathrm{T}}\mathbf{u}^\star + \operatorname{sign}\left(\mathbf{x}_j^\star\right)\gamma_{r(\mathbf{x}^\star,j)} \geq \tfrac{1}{n} > 0 \tag{20}$$

which is in contradiction with (B-10a).

Similarly to Counterexample 1, Counterexample 2 describes a simple setup where both $\mathbf{x}^\star$ and $\mathbf{u}^\star$ are unique and can

be exhibited in closed-form. However, it appears that the latter couple of primal-dual solutions does not satisfies the optimality conditions derived in [2] for the OWL problem.

Prior to conclude our note, we provide some elements regarding the fallacy in the derivation of the optimality conditions (B-10a) and (B-10b). An attentive reading of [2, Section 3.1] suggests that Bao *et al.* implicitly assume that addressing the following optimization problem (see [2, Eq. (8)])

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{u}^\star + \sum_{k=1}^{n} \gamma_k |\mathbf{x}|_{[k]} \tag{B-21}$$

is equivalent to solve

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{u}^\star + \sum_{j=1}^{n} \gamma_{\mathrm{r}(\mathbf{x}^\star,j)} |\mathbf{x}_j|, \tag{B-22}$$

where $\mathbf{x}^\star$ denotes a minimizer of (1 - OWL). In other words, the authors assume that if the order of the entries (in absolute value) of the solution of (B-21) is known, then one can equivalently solve a simpler optimization problem that already takes into account the permutation of the entries. Besides, the authors did not mention how to handle the design of the rank function in presence of multiple minimizers.

To support our disagreement with this statement, consider again the setup described in Counterexample 2. Since all entries of $\mathbf{x}^\star$ are equal (and positive), $[1,\dots,n]$ is a valid ordering. Yet, solving (B-21) is not equivalent to solving

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{u}^\star + \gamma_1 |\mathbf{x}_1| \tag{B-23}$$

which admits $-\infty$ as minimum.

APPENDIX A
PROOF RELATED TO COUNTEREXAMPLE 1

In this appendix, we exhibit the (unique) solution of the OWL problem from Counterexample 1. Our proof is organized as follows. We first state in Section A-A a technical lemma. We show in Section A-B that, in this example, all minimizers of OWL have nonnegative entries. Then, we demonstrate in Section A-C that all minimizers are proportional to the "all-one" vector. We finally solve in Section A-D the OWL problem and show that the minimizer is in fact unique.

In the following, we let $\mathbf{y}\in\mathbb{R}^m$, $\mathbf{A}\in\mathbb{R}^{m\times n}$ and $\{\gamma_k\}_{k=1}^n$ be defined as in Counterexample 1.

A. A technical lemma

The following results will be used in our subsequent derivations:

Lemma 1. *Let $\mathbf{u}^\star\in\mathbb{R}^m$ be the (unique) maximizer of the dual problem of OWL. Then, in the setup of Counterexample 1, we have:*

$$\mathbf{1}_m^{\mathrm{T}}\mathbf{u}^\star \geq 0. \tag{24}$$

Proof. See first that the (primal) objective function P rewrites $\forall \mathbf{x}\in\mathbb{R}^n$ as $P(\mathbf{x}) = f(\mathbf{A}\mathbf{x}) + r_{\mathrm{OWL}}(\mathbf{x})$ and recall that r_{OWL} defines a norm under (H-3), see [4, Proposition 1.1]. Hence, its Fenchel dual problem writes [1, Eq. (1.4)]

$$\begin{aligned} \underset{\mathbf{u}\in\mathbb{R}^m}{\arg\max} \ D(\mathbf{u}) = f^*(\mathbf{u}) &= \tfrac{1}{2}\|\mathbf{y}\|_2^2 - \tfrac{1}{2}\|\mathbf{y}-\mathbf{u}\|_2^2 \\ \text{s.t. } & r_{\mathrm{OWL}}^*(\mathbf{A}^{\mathrm{T}}\mathbf{u}) \leq 1 \end{aligned} \tag{25}$$

where f^* denotes the Fenchel conjugate function of f and r_{OWL}^* the dual norm of r_{OWL}.

Since r_{OWL}^* is also a norm, we have $r_{\mathrm{OWL}}^*(\mathbf{A}^{\mathrm{T}}\mathbf{0}_m) = 0 \leq 1$ so that the vector of zeros $\mathbf{0}_m$ is admissible. Using the fact that f^* is concave and that $\nabla f^*(\mathbf{0}_m) = \mathbf{y}$, we have for all $\mathbf{u}\in\mathbb{R}^m$:

$$f^*(\mathbf{u}) \leq f^*(\mathbf{0}_m) + \mathbf{y}^{\mathrm{T}}\mathbf{u}. \tag{26}$$

Recalling that $\mathbf{y} = \mathbf{1}_m$ and particularizing the latter inequality to $\mathbf{u} = \mathbf{u}^\star$, we obtain

$$\mathbf{1}_m^{\mathrm{T}}\mathbf{u}^\star \geq f^*(\mathbf{u}^\star) - f^*(\mathbf{0}_m) \geq 0 \tag{27}$$

since $\mathbf{u}^\star$ is the maximizer of (25). This concludes the proof. □

B. All minimizers have positive entries

Let $\mathbf{x}^\star\in\mathbb{R}^n$ be a minimizer of OWL. In the following, we show by contradiction that $\mathbf{x}_j^\star \geq 0$ for all $j\in[\![1,n]\!]$.

Assume that there exists $j_0\in[\![1,n]\!]$ such that $\mathbf{x}_{j_0}^\star < 0$. Suppose moreover (without loss of generality and up to a permutation of the indices) that the entries of $\mathbf{x}^\star$ are ordered (in absolute value), *i.e.*, that

$$|\mathbf{x}_1^\star| \geq |\mathbf{x}_2^\star| \geq \cdots \geq |\mathbf{x}_n^\star| \tag{28}$$

and that

$$\left|\mathbf{x}_{j_0}^\star\right| > \left|\mathbf{x}_{j_0+1}^\star\right| \tag{29}$$

with the convention $\mathbf{x}_{n+1}^\star = 0$. Define $\mathbf{x}'\in\mathbb{R}^n$ as

$$\mathbf{x}' = \mathbf{x}^\star + \varepsilon_1 \mathbf{e}_{j_0} \tag{30}$$

where $\mathbf{e}_{j_0}$ denote the j_0th vector of the canonical basis of $\mathbb{R}^n$ and ε_1 satisfies

$$0 < \varepsilon_1 < \min\left(\left|\mathbf{x}_{j_0}^\star\right| - \left|\mathbf{x}_{j_0+1}^\star\right|, \frac{2\gamma_{j_0}}{n}\right). \tag{31}$$

Note that such ε_1 exists since $\gamma_{j_0} > 0$ by hypothesis and that $\left|\mathbf{x}_{j_0}^\star\right| > \left|\mathbf{x}_{j_0+1}^\star\right|$. Moreover, the ordering of $\mathbf{x}'$ is also conserved, that is

$$|\mathbf{x}_1'| \geq |\mathbf{x}_2'| \geq \cdots \geq |\mathbf{x}_n'| \tag{32}$$

and $\left|\mathbf{x}_{j_0}'\right| = \left|\mathbf{x}_{j_0}^\star\right| - \varepsilon_1$ by design of ε_1. Hence,

$$r_{\mathrm{OWL}}(\mathbf{x}') = r_{\mathrm{OWL}}(\mathbf{x}^\star) - \gamma_{j_0}\varepsilon_1. \tag{33}$$

Using the fact that $\mathbf{A}\mathbf{e}_{j_0} = \mathbf{a}_{j_0} = \mathbf{1}_m$ by definition of $\mathbf{A}$, we have

$$\begin{aligned} P(\mathbf{x}') &= P(\mathbf{x}^\star) + \frac{\varepsilon_1^2}{2}\|\mathbf{1}_m\|_2^2 - \varepsilon_1 \mathbf{1}_m^{\mathrm{T}}(\mathbf{y}-\mathbf{A}\mathbf{x}^\star) - \gamma_{j_0}\varepsilon_1 \\ &\overset{(6)}{=} P(\mathbf{x}^\star) + \frac{\varepsilon_1^2}{2}\|\mathbf{1}_m\|_2^2 - \varepsilon_1 \mathbf{1}_m^{\mathrm{T}}\mathbf{u}^\star - \gamma_{j_0}\varepsilon_1 \\ &\overset{\text{Lemma 1}}{\leq} P(\mathbf{x}^\star) + \varepsilon_1\left(\frac{n}{2}\varepsilon_1 - \gamma_{j_0}\right). \end{aligned} \tag{34}$$

By choice of ε_1 (see (31)) we have

$$\varepsilon_1\Big(\frac{n}{2}\varepsilon_1 - \gamma_{j_0}\Big) < 0 \tag{35}$$

so that $P(\mathbf{x}') < P(\mathbf{x}^\star)$ which is the desired contradiction. Therefore, the entries of all minimizers of OWL have nonnegative entries.

C. All minimizers are proportional to $\mathbf{1}_n$

We now show that all the solutions of OWL are proportional to vector $\mathbf{1}_n$. We again proceed by contradiction: let $\mathbf{x}^\star$ be a minimizer of (1 - OWL) and assume that there exists $j_0 \in [\![1, n-1]\!]$ such that $\mathbf{x}^\star_{j_0} > \mathbf{x}^\star_{j_0+1}$ (recall that all entries of $\mathbf{x}^\star$ are nonnegative). Assume moreover and without loss of generality (up to a permutations of the columns of $\mathbf{A}$) that

$$\mathbf{x}^\star_1 \geq \mathbf{x}^\star_2 \geq \cdots \geq \mathbf{x}^\star_n. \tag{36}$$

We follow a similar rationale as in Section A-B: define the vector $\mathbf{x}' \in \mathbb{R}^n$ such that

$$\mathbf{x}' = \mathbf{x}^\star - \varepsilon_2 \mathbf{e}_{j_0} + \varepsilon_2 \mathbf{e}_{j_0+1} \tag{37}$$

where $\mathbf{e}_{j_0}, \mathbf{e}_{j_0+1}$ denote the j_0, j_0+1th vectors of the canonical basis of $\mathbb{R}^n$ and ε_2 is a scalar such that

$$0 < \varepsilon_2 < \frac{\mathbf{x}^\star_{j_0} - \mathbf{x}^\star_{j_0+1}}{2}. \tag{38}$$

Note that such a scalar exists since $\mathbf{x}^\star_{j_0} > \mathbf{x}^\star_{j_0+1}$ by assumption. See also that $\mathbf{x}'$ is nonnegative by definition, and that

$$\mathbf{x}'_1 \geq \mathbf{x}'_2 \geq \cdots \geq \mathbf{x}'_n \geq 0 \tag{39}$$

Hence,

$$r_{\text{OWL}}(\mathbf{x}') = r_{\text{OWL}}(\mathbf{x}^\star) + \varepsilon_2(\gamma_{j_0+1} - \gamma_{j_0}) < r_{\text{OWL}}(\mathbf{x}^\star) \tag{40}$$

by definition of ε_2 and the sequence of weights $\{\gamma_k\}_{k=1}^n$.

Recalling that $\mathbf{A}\mathbf{e}_j = \mathbf{1}_m$ for all $j \in [\![1, n]\!]$, we have

$$\begin{aligned} P(\mathbf{x}') &= \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}^\star\|_2^2 + r_{\text{OWL}}(\mathbf{x}') \\ &\overset{(40)}{<} \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}^\star\|_2^2 + r_{\text{OWL}}(\mathbf{x}^\star) \\ &= P(\mathbf{x}^\star) \end{aligned} \tag{41}$$

which leads to the desired contradiction. Hence all solutions of OWL are proportional to the vector of ones.

D. Solving the OWL problem

Combining the results of Sections A-B and A-C, we have that solving (1 - OWL) is equivalent[2] to solve

$$\operatorname*{arg\,min}_{x \in \mathbb{R}_+} \ \tilde{P}(x) \triangleq \tfrac{1}{2}\|\mathbf{y} - x\mathbf{A}\mathbf{1}_n\|_2^2 + x\sum_{k=1}^n \gamma_k. \tag{42}$$

We let the reader check that the scalar $x^\star$ defined in (13) satisfies $\tilde{P}'(x^\star) = 0$. Moreover, one sees that $x^\star > 0$ using the fact that $\gamma_1 \leq 1$ and (H-3). Therefore, $x^\star$ satisfies the optimality condition of (the strictly-convex) optimization problem (42) and is therefore its unique minimizer.

[2] In the sense that there exists a bijection between the set of minimizers.

APPENDIX B
PROOF RELATED TO COUNTEREXAMPLE 2

Let $\mathbf{y} \in \mathbb{R}^m, \mathbf{A} \in \mathbb{R}^{m\times n}$ and $\{\gamma_k\}_{k=1}^n$ be defined as in Counterexample 2. We first note that, for such choices of the parameters, OWL reduces to (17) since r_{OWL} simplifies as

$$\forall \mathbf{x} \in \mathbb{R}^n, \quad r_{\text{OWL}}(\mathbf{x}) = |\mathbf{x}|_{[1]} = \max_j(\{|\mathbf{x}_j|\}) = \|\mathbf{x}\|_\infty. \tag{43}$$

Note also that the solution of (17) exists and is unique because the function to minimize is continuous, strongly-convex and $\mathbb{R}^n$ is a nonempty, convex and closed set.

Let $\mathbf{x}^\star = (n-1)/n\mathbf{1}_n$. Since the function to minimize is lower-semicontinuous, a direct application of Fermat's rule [3, Theorem 16.3] to problem (45) shows that $\mathbf{x}^\star$ is a minimizer of (17) if and only if

$$\mathbf{y} - \mathbf{x}^\star = \tfrac{1}{n}\mathbf{1}_n \in \partial\|\mathbf{x}^\star\|_\infty. \tag{44}$$

Since the function $\mathbf{x} \to \|\mathbf{x}\|_\infty$ defines a norm, its subdifferential is well defined for all $\mathbf{x} \in \mathbb{R}^n$ and writes [1, Eq. (1.4)]:

$$\partial\|\mathbf{x}\|_\infty = \big\{\mathbf{g} \in \mathbb{R}^n \colon \|\mathbf{g}\|_1 \leq 1 \text{ and } \mathbf{x}^{\mathrm{T}}\mathbf{g} = \|\mathbf{x}\|_\infty\big\}. \tag{45}$$

Let us show now that $\mathbf{y} - \mathbf{x}^\star \in \partial\|\mathbf{x}^\star\|_\infty$. Since $\mathbf{y} - \mathbf{x}^\star$ has nonnegative entries, we first have

$$\|\mathbf{y} - \mathbf{x}^\star\|_1 = \tfrac{1}{n}\|\mathbf{1}_n\|_1 = 1. \tag{46}$$

See then that

$$\mathbf{x}^{\star\mathrm{T}}(\mathbf{y} - \mathbf{x}^\star) = \frac{n-1}{n} = \|\mathbf{x}^\star\|_\infty. \tag{47}$$

Hence, $\mathbf{y} - \mathbf{x}^\star \in \partial\|\mathbf{x}^\star\|_\infty$ which concludes the proof.

REFERENCES

[1] F. BACH, R. JENATTON, J. MAIRAL, AND G. OBOZINSKI, *Convex optimization with sparsity-inducing norms*, in Optimization for Machine Learning, Neural information processing series, MIT Press, 2011, pp. 19–49, https://doi.org/10.7551/mitpress/8996.003.0004.

[2] R. BAO, B. GU, AND H. HUANG, *Fast OSCAR and OWL with safe screening rules*, in Proceedings of the 37th International Conference on Machine Learning, 2020.

[3] H. H. BAUSCHKE AND P. L. COMBETTES, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer International Publishing, 2017, https://doi.org/10.1007/978-3-319-48311-5.

[4] M. BOGDAN, E. VAN DEN BERG, W. SU, AND E. CANDES, *Statistical estimation and testing via the sorted l1 norm*, 2013, https://arxiv.org/abs/1310.1969.

[5] L. GHAOUI, V. VIALLON, AND T. RABBANI, *Safe feature elimination in sparse supervised learning*, Pacific Journal of Optimization, 8 (2010).

[6] E. NDIAYE, O. FERCOQ, ALEX, RE GRAMFORT, AND J. SALMON, *Gap safe screening rules for sparsity enforcing penalties*, Journal of Machine Learning Research, 18 (2017), pp. 1–33, http://jmlr.org/papers/v18/16-577.html.