Expanding boundaries of Gap Safe screening

Cassio F. Dantas, Emmanuel Soubies, Cédric Févotte

Journal of Machine Learning Research 1 (2021) 1-56 Submitted 02/21; Published 00/00

Expanding Boundaries of Gap Safe Screening

Cassio F. Dantas CASSIO.FRAGA-DANTAS@IRIT.FR
Emmanuel Soubies EMMANUEL.SOUBIES@IRIT.FR
Cédric Févotte CEDRIC.FEVOTTE@IRIT.FR
IRIT, Université de Toulouse, CNRS, Toulouse, France

Editor: –

Abstract

Sparse optimization problems are ubiquitous in many fields such as statistics, signal/image processing and machine learning. This has led to the birth of many iterative algorithms to solve them. A powerful strategy to boost the performance of these algorithms is known as *safe screening*: it allows the early identification of zero coordinates in the solution, which can then be eliminated to reduce the problem's size and accelerate convergence. In this work, we extend the existing Gap Safe screening framework by relaxing the global strong-concavity assumption on the dual cost function. Instead, we exploit local regularity properties, that is, strong concavity on well-chosen subsets of the domain. The non-negativity constraint is also integrated to the existing framework. Besides making safe screening possible to a broader class of functions that includes β-divergences (e.g., the Kullback-Leibler divergence), the proposed approach also improves upon the existing Gap Safe screening rules on previously applicable cases (e.g., logistic regression). The proposed general framework is exemplified by some notable particular cases: logistic function, $\beta = 1.5$ and Kullback-Leibler divergences. Finally, we showcase the effectiveness of the proposed screening rules with different solvers (coordinate descent, multiplicative-update and proximal gradient algorithms) and different data sets (binary classification, hyperspectral and count data).

Keywords: Convex optimization, safe screening rules, sparse regression, β-divergence, non-negativity

1. Introduction

Safe screening rules have proved to be very powerful tools in order to accelerate the resolution of large-scale sparse optimization problems that arise in statistics, machine learning, signal/image inverse problems, pattern recognition, among other fields. The very principle of safe screening is to identify the zero coordinates in the solution before and/or within the course of iterations of any solver. Once identified, these inactive coordinates can be screened out, thus reducing the size of the problem and consequently the computational load of the solver. Hence, should a screening rule allow to screen many coordinates with a low computational overhead, significant speedups can be observed in practice– see for instance El Ghaoui et al. (2012); Bonnefoy et al. (2015); Ndiaye et al. (2017), and Section 5.

A brief tour of existing screening strategies. Safe screening rules were initially proposed for the Lasso problem (El Ghaoui et al., 2012) and were later extended to some of its variants: group Lasso (Wang et al., 2015a; Bonnefoy et al., 2015), sparse group Lasso (Ndiaye et al., 2016; Wang et al., 2019), non-negative Lasso (Wang et al., 2019), fused Lasso (Wang

et al., 2015), and generalized Lasso (Ren et al., 2018). As opposed to correlation-based feature selection techniques (Fan and Lv, 2008; Tibshirani et al., 2011), safe screening strategies are guaranteed to remove only coordinates that do no belong to the solution support. Beyond Lasso, safe screening has also been used for other machine learning problems, such as: binary logistic regression (El Ghaoui et al., 2012; Ndiaye et al., 2017), metric learning (Yoshida et al., 2018), nuclear norm minimization (Zhou and Zhao, 2015) and support vector machine (Ogawa et al., 2013; Wang et al., 2014; Zimmert et al., 2015).

Three main classes of screening rules can be distinguished: 1) Static rules (El Ghaoui et al., 2012; Xiang et al., 2011; Xiang and Ramadge, 2012) perform variable elimination once and for all prior to the optimization process; 2) Dynamic rules (Bonnefoy et al., 2015; Ndiaye et al., 2017) perform screening repeatedly over the iterations of an iterative solver, leveraging the improvement of the solution estimate to screen-out more coordinates; 3) Sequential rules (Xiang et al., 2011; Wang et al., 2015a; Liu et al., 2014; Malti and Herzet, 2016) exploit information from previously-solved problems in a regularization path approach. See, for instance, Xiang et al. (2017) or Ndiaye (2018) for a survey of the domain.

Most of the mentioned screening techniques are problem-specific, as they exploit particular properties of the targeted loss function. For instance, rules in El Ghaoui et al. (2012); Bonnefoy et al. (2015); Wang et al. (2015a) assume the dual problem to be a projection problem, which is no longer the case for non-quadratic loss functions. The Gap Safe rule (Ndiaye et al., 2017), however, relies primarily on the duality gap which is defined for any primal-dual pair of problems, regardless of the specific cost functions. The authors were therefore able to deploy this screening rule for a fairly generic class of functions. Additionally, this particular rule leads to state-of-the-art performances in a wide range of scenarios (Ndiaye et al., 2017).

Problem definition and working assumptions. In this work, we consider the following generic primal problem

$$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{C}} P_\lambda(\mathbf{x}) := F(\mathbf{A}\mathbf{x}) + \lambda\Omega(\mathbf{x}) \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $F : \mathbb{R}^m \to \mathbb{R}$, $\Omega : \mathbb{R}^n \to \mathbb{R}_+$, $\mathcal{C} \subseteq \mathbb{R}^n$, and $\lambda > 0$. Moreover, we make the following assumptions:

- F is coordinate-wise separable, i.e., $F(\mathbf{z}) = \sum_{i=1}^m f_i(z_i)$ where each scalar function $f_i : \mathbb{R} \to \mathbb{R}$ is proper, lower semi-continuous, convex, and differentiable.
- Ω is a group-decomposable norm, i.e., given a partition $\mathcal{G}$ of $\{1,\dots,n\}$, $\Omega(\mathbf{x}) = \sum_{g\in\mathcal{G}} \Omega_g(\mathbf{x}_g)$ where each Ω_g is a norm on $\mathbb{R}^{n_g}$ (n_g denoting the cardinality of $g \in \mathcal{G}$).[1]
- $\mathcal{C}$ is a constraint set. Here, we study the cases $\mathcal{C} = \mathbb{R}^n$ (unconstrained) and $\mathcal{C} = \mathbb{R}^n_+$ (non-negativity constraint).

Finally, we assume that P_λ admits at least a minimizer $\mathbf{x}^\star \in \mathcal{C}$.

1. The ℓ_1-norm is a trivial example of group-decomposable norm, where each group g corresponds to a singleton (i.e., $\mathcal{G} = \{\{1\}, \ \dots \{n\}\}$) and $\Omega_g = |\cdot|$.

Contributions and roadmap. The present paper extends the Gap Safe rules proposed in Ndiaye et al. (2017) (and recalled in Section 2) to a broader class of problems of the form (1) in two aspects. First, we allow the use of a non-negativity constraint ($\mathcal{C} = \mathbb{R}^n_+$). Second, we relax the requirement of global strong concavity of the dual objective function (see Section 3). Indeed, we prove in Theorem 5 that a Gap Safe sphere can be constructed from the only requirement that the dual objective function is locally strongly concave on a subset that contains the dual solution. This result is exploited in Section 3.1 to revisit the Gap Safe dynamic screening algorithm (Ndiaye et al., 2017). It allows to tackle problems such as common ℓ_1-regularized Kullback-Leibler regression. In Section 3.2, we further exploit Theorem 5 to propose a new Gap Safe dynamic screening algorithm where, at each iteration, the Gap Safe sphere is iteratively refined, leading to an increase of the number of screened variables. Finally, these two generic approaches are applied to a set of concrete problems in Section 4, and are experimentally evaluated in Section 5.

2. Safe Screening for Generalized Linear Models

2.1 Notations and Definitions

Scalar operations (such as division, logarithm, exponential and comparisons), whenever applied to vectors, are implicitly assumed as entry-wise operations. We denote by $[n] = \{1, \dots, n\}$ the set of integers ranging from 1 to $n \in \mathbb{N}$. For a vector $\mathbf{z} \in \mathbb{R}^n$, z_i (or sometimes $[\mathbf{z}]_i$ to avoid ambiguities) stands for its i-th entry. We use the notation $[\mathbf{z}]^+$ to refer to the positive rectification (ReLU) operation defined as $\max(0, z_i)$ for all $i \in [n]$. Given a subset of indices $g \subseteq [n]$ with cardinality $|g| = n_g$, $\mathbf{z}_g \in \mathbb{R}^{n_g}$ (in bold case) denotes the restriction of $\mathbf{z}$ to its entries indexed by the elements of g. For a matrix $\mathbf{A}$, we denote by $\mathbf{a}_j$ its j-th column and $\mathbf{A}_g$ the matrix formed out of the columns of $\mathbf{A}$ indexed by the set $g \subseteq [n]$. We denote $\mathcal{I}^\complement$ the complement of a set $\mathcal{I}$.

We consider functions taking values over the extended real line where the *domain* of a function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is defined as the set:

$$\operatorname{dom}(f) = \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) < +\infty\}.$$

For a function $f : \mathbb{R}^n \to (-\infty, +\infty]$, we denote $f^* : \mathbb{R}^n \to (-\infty, +\infty]$ its *Fenchel-Legendre transform* (or conjugate function), defined as follows:

$$f^*(\mathbf{u}) := \sup_{\mathbf{z} \in \mathbb{R}^n} \langle \mathbf{z}, \mathbf{u} \rangle - f(\mathbf{z}). \tag{2}$$

For a norm Ω over $\mathbb{R}^n$, we denote $\overline{\Omega}$ its associated dual norm such that:

$$\overline{\Omega}(\mathbf{u}) := \sup_{\Omega(\mathbf{z}) \leq 1} \langle \mathbf{z}, \mathbf{u} \rangle. \tag{3}$$

The dual norm $\overline{\Omega}$ is not to be confused with the Fenchel conjugate Ω^*. Actually, the conjugate of a norm is the indicator function of the unit ball of the dual norm, i.e., $\Omega^*(\mathbf{u}) = \mathbb{1}_{\overline{\Omega}(\mathbf{u}) \leq 1}$.

2.2 Dual Problem

In Theorem 1 below, we derive the dual problem of (1) together with the associated primal-dual optimality conditions. In particular, we provide a generic expression for the two

considered constraint sets $\mathcal{C} = \mathbb{R}^n$ and $\mathcal{C} = \mathbb{R}^n_+$ which extends Ndiaye et al. (2017, Theorem 2). The proof is given in Appendix A.

Theorem 1. *The dual formulation of the optimization problem defined in* (1) *is given by*

$$\boldsymbol{\theta}^\star \in \underset{\boldsymbol{\theta}\in\Delta_{\mathbf{A}}}{\operatorname{argmax}} \; D_\lambda(\boldsymbol{\theta}) := -\sum_{i=1}^{n} f_i^*(-\lambda\theta_i) \tag{4}$$

$$\text{with} \;\; \Delta_{\mathbf{A}} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \forall g \in \mathcal{G}, \overline{\Omega}_g(\phi(\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta})) \leq 1\} \cap \operatorname{dom}(D_\lambda) \tag{5}$$

where $\boldsymbol{\theta} \in \mathbb{R}^m$ *is the dual variable and "* $\leq$ *" is defined component-wisely. The function* $\phi : \mathbb{R} \to \mathbb{R}$

$$\phi(x) = \begin{cases} x, & \text{if } \; \mathcal{C} = \mathbb{R}^m \\ [x]^+, & \text{if } \; \mathcal{C} = \mathbb{R}^m_+ \end{cases} \tag{6}$$

is also applied component-wisely in (5)*. Moreover, the first-order optimality conditions for a primal-dual solution pair* $(\mathbf{x}^\star, \boldsymbol{\theta}^\star) \in (\operatorname{dom}(P_\lambda) \cap \mathcal{C}) \times \Delta_{\mathbf{A}}$*, are given by*

$$\forall i \in [m], \quad \lambda\theta_i^\star = -f_i'([\mathbf{A}\mathbf{x}^\star]_i) \qquad \textit{(primal-dual link)} \tag{7}$$

$$\forall g \in \mathcal{G}, \; \begin{cases} \overline{\Omega}_g(\phi(\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta}^\star)) \leq 1, & \text{if } \; \mathbf{x}_g^\star = \mathbf{0} \\ \overline{\Omega}_g(\phi(\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta}^\star)) = 1, \quad (\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta}^\star)^\mathsf{T}\mathbf{x}_g^\star = \Omega_g(\mathbf{x}_g^\star), & \text{otherwise} \end{cases} \qquad \textit{(sub-differential inclusion)} \tag{8}$$

2.3 Safe Screening Rules

A direct consequence of Theorem 1 is that, given the dual solution $\boldsymbol{\theta}^\star$,

$$\overline{\Omega}_g(\phi(\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta}^\star)) < 1 \implies \mathbf{x}_g^\star = \mathbf{0} \tag{9}$$

for any primal solution $\mathbf{x}^\star$. Hence, every group of coordinates $g \in \mathcal{G}$ for which $\overline{\Omega}_g(\phi(\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta}^\star)) < 1$ is surely inactive (i.e., $\mathbf{x}_g = \mathbf{0}$). They can thus be safely screened out in order to reduce the size of the primal problem and accelerate its resolution. In practice, however, the dual solution $\boldsymbol{\theta}^\star$ is unknown (in advance) and (9) cannot be evaluated. Fortunately, it is possible to define a more restrictive—yet practical—sufficient condition that relies on the concept of *safe region*.

Definition 2 (Safe Region)**.** *A compact subset* $\mathcal{R} \subset \mathbb{R}^n$ *is said to be a* safe region *if it contains the dual solution (i.e.,* $\boldsymbol{\theta}^\star \in \mathcal{R}$*).*

Proposition 3 (Safe Screening Rule (El Ghaoui et al., 2012))**.** *Let* $\mathcal{R}$ *be a safe region and* $g \in \mathcal{G}$*. Then,*

$$\max_{\boldsymbol{\theta}\in\mathcal{R}} \overline{\Omega}_g(\phi(\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta})) < 1 \implies \overline{\Omega}_g(\phi(\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta}^\star)) < 1 \implies \mathbf{x}_g^\star = \mathbf{0} \tag{10}$$

The quantity $\max_{\boldsymbol{\theta}\in\mathcal{R}} \overline{\Omega}_g(\phi(\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta}))$ is referred to as *screening test* as it allows to test whether a group of coordinates is guaranteed to be zero in the optimal solution.

With Proposition 3, numerous screening rules can be defined from the construction of different safe regions. Although any region $\mathcal{R}$ such that $\Delta_{\mathbf{A}} \subset \mathcal{R}$ is safe as $\boldsymbol{\theta}^\star \in \Delta_{\mathbf{A}}$, these trivial choices would lead to poor screening performance. Instead, to maximise the number of screened groups while limiting the computational overhead of testing, one needs to construct safe regions $\mathcal{R}$ that are as small as possible, and for which the quantity $\max_{\boldsymbol{\theta}\in\mathcal{R}} \overline{\Omega}_g(\phi(\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta}))$ (screening test) can be computed efficiently. It is thus standard practice to consider simple regions such as balls (El Ghaoui et al., 2012; Bonnefoy et al., 2015; Ndiaye et al., 2017) or domes (Fercoq et al., 2015; Xiang and Ramadge, 2012) as they are more likely to lead to closed-form expressions of the screening test (see Section 4).

Gap Safe Sphere. A notable safe region is the Gap Safe sphere as it leads to state-of-the-art screening performances in a wide range of scenarios (Ndiaye et al., 2017). It relies on the duality gap for the primal-dual problems (1)-(4) defined as

$$\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta}) := P_\lambda(\mathbf{x}) - D_\lambda(\boldsymbol{\theta}). \tag{11}$$

Theorem 4 (Gap Safe Sphere (Ndiaye et al., 2017))**.** *Assuming that the dual function D_λ is α-strongly concave, then for any feasible primal-dual pair $(\mathbf{x}, \boldsymbol{\theta}) \in (\mathrm{dom}(P_\lambda) \cap \mathcal{C}) \times \Delta_{\mathbf{A}}$:*

$$\mathcal{B}(\boldsymbol{\theta}, r), \ \textit{with } r = \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha}} \tag{12}$$

is a safe region, i.e., $\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r)$.

The Gap Safe sphere improves over previously proposed safe regions (El Ghaoui et al., 2012; Bonnefoy et al., 2015; Wang et al., 2015b) in two ways. First, it is not restricted to the Lasso problem and applies to a broad class of problems of the form (1), under the assumption that the associated dual function D_λ is strongly concave. Second, in case of strong duality, its radius vanishes when a converging sequence of primal-dual variables is provided (with the duality gap tending to zero).

2.4 Screening with Existing Solvers

The previously presented screening tools can be integrated to most existing solvers in order to reduce the size of the primal problem (1) and accelerate its resolution. As previously mentioned, screening rules can be exploited in many ways (see Xiang et al., 2017) that include *static screening* (El Ghaoui et al., 2012), *sequential screening* (Xiang et al., 2011), or *dynamic screening* (Bonnefoy et al., 2015). In this work, we focus on the dynamic screening approach that fully exploits the structure of iterative optimization algorithms by screening out groups of coordinates in the course of iterations. As the algorithm converges, smaller safe regions can be defined, leading to an increasing number of screened groups. More precisely, we use the Gap Safe dynamic screening scheme proposed by Ndiaye et al. (2017) as our baseline. This scheme is proposed in Algorithm 1 where

$$\{\mathbf{x}, \boldsymbol{\eta}\} \leftarrow \texttt{PrimalUpdate}(\mathbf{x}, \mathbf{A}, \lambda, \boldsymbol{\eta}) \tag{13}$$

represents the update step of any iterative primal solver for (1). There, $\mathbf{x}$ denotes the primal variable and $\boldsymbol{\eta}$ is a vector formed out of the auxiliary variables of the solver (e.g.,

Algorithm 1 Dynamic Gap Safe Screening (DGS) (Ndiaye et al., 2017):
$\hat{\mathbf{x}} = \texttt{GAPSolver}(\mathbf{A}, \lambda, \varepsilon_{\text{gap}})$

1: **Initialize** $\mathcal{A} = \mathcal{G}$, $\mathbf{x} \in \mathcal{C}$
2: **Set** $\boldsymbol{\eta}$ according to the solver
3: **Compute** α a strong concavity bound of D_λ on $\mathbb{R}^m$
4: **repeat**
5: — *Solver update restricted to preserved set* —
6: $\{\mathbf{x}_\mathcal{A}, \boldsymbol{\eta}\} \leftarrow \texttt{PrimalUpdate}(\mathbf{x}_\mathcal{A}, \mathbf{A}_\mathcal{A}, \lambda, \boldsymbol{\eta})$
7: — *Dynamic Screening* —
8: $\boldsymbol{\theta} \leftarrow \boldsymbol{\Theta}(\mathbf{x}) \in \Delta_\mathbf{A}$ ▷ *Dual update*
9: $r \leftarrow \sqrt{\frac{2\,\text{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha}}$ ▷ *Safe radius*
10: $\mathcal{A} \leftarrow \{g \in \mathcal{A} \mid \max_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}, r)} \overline{\Omega}_g(\phi(\mathbf{A}_g^\mathsf{T} \boldsymbol{\theta})) \geq 1\}$ ▷ *Screening test*
11: $\mathbf{x}_{\mathcal{A}^c} \leftarrow \mathbf{0}$
12: **until** $\text{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta}) < \varepsilon_{\text{gap}}$

gradient step-size, previous primal estimates). To keep the presentation concise, screening is performed after every iteration of the primal solver in Algorithm 1. However, it is noteworthy to mention that this is not a requirement. Screening can actually be performed at any chosen moment. For instance, on regular intervals between a certain number of iterations of the solver. Finally, let us emphasise the nested update of the preserved set (line 10) showing that the screened groups are no longer tested in the ensuing iterations.

To construct a Gap Safe sphere, a dual feasible point $\boldsymbol{\theta} \in \Delta_\mathbf{A}$ is required (Theorem 4). Although such a dual point comes for free when deploying primal-dual solvers (Chambolle and Pock, 2011; Yanez and Bach, 2017) for (1), it needs to be computed from $\mathbf{x}$ when the solver only provides a primal solution estimate at each iteration. Needless to say that the latter is the case of many popular solvers for (1) such as (Beck and Teboulle, 2009; Harmany et al., 2012; Hsieh and Dhillon, 2011). At line 8 of Algorithm 1, this computation of a dual feasible point (referred to as dual update) is defined through the function $\boldsymbol{\Theta} : \mathbb{R}^n \to \Delta_\mathbf{A}$. Always with the aim of maximizing the number of screened groups (i.e., reducing the Gap Safe sphere) while limiting the computational overhead, a rule of thumb for $\boldsymbol{\Theta}$ is that, for a primal estimate $\mathbf{x}$, the evaluation of $\boldsymbol{\Theta}(\mathbf{x})$ only requires "simple" operations and $\|\boldsymbol{\Theta}(\mathbf{x}) - \boldsymbol{\theta}^\star\|$ is as small as possible. By exploiting the optimality condition (7), it is customary to define $\boldsymbol{\Theta}$ as a simple rescaling of $\nabla F(\mathbf{A}\mathbf{x})$ (see Section 4.1.4). Not only this choice is computationally cheap but it enjoys the appealing property that $\boldsymbol{\Theta}(\mathbf{x}) \to \boldsymbol{\theta}^\star$ as $\mathbf{x} \to \mathbf{x}^\star$.

The purpose of the next section is to extend and refine Algorithm 1.

3. Exploiting Local Regularity Properties of the Dual Function

The Gap Safe sphere given in Theorem 4 requires the dual function to be α-strongly concave on $\mathbb{R}^m$. This precludes its application to an important class of problems of practical interest. For instance, problems involving the Kullback-Leibler divergence and other β-divergences with $\beta \in (1, 2)$ (see Section 4). In this section, we relax this hypothesis by leveraging only *local* properties of the dual function. More precisely, we derive in Theorem 5 a Gap Safe

sphere with the only requirement that D_λ is strongly concave on a well-chosen subset of its domain. The proof is provided in Appendix B.

Theorem 5. *Assume that D_λ is $\alpha_\mathcal{S}$-strongly concave on a subset $\mathcal{S} \subset \mathbb{R}^m$ such that $\boldsymbol{\theta}^\star \in \mathcal{S}$. Then, for any feasible primal-dual pair $(\mathbf{x}, \boldsymbol{\theta}) \in (\mathrm{dom}(P_\lambda) \cap \mathcal{C}) \times (\Delta_\mathbf{A} \cap \mathcal{S})$:*

$$\mathcal{B}(\boldsymbol{\theta}, r), \text{ with } r = \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha_\mathcal{S}}} \tag{14}$$

is a Gap Safe sphere, i.e., $\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r)$.

Theorem 5 allows us to extend the application of Gap Safe rules to problems for which the corresponding dual function D_λ is not globally strongly concave. When it comes to the primal objective function, it allows us to tackle some data-fidelity functions which do not have a Lipschitz-continuous gradient.

Moreover, this result can also be used to improve upon the performance of standard Gap Safe rules by providing better local bounds $\alpha_\mathcal{S}$ for the strong concavity of the dual function. Indeed, a global strong concavity bound α (if any) cannot ever be larger than its local counterpart α_S for a given valid set $\mathcal{S}$. Hence, not only Theorem 5 extends Gap Safe rules to a broader class of problems, but it can boost their performances when the known global strong concavity bound is poor (too small).

In the two following sections, we exploit Theorem 5 to revisit (in Section 3.1) and improve (in Section 3.2) the Gap Safe screening approach proposed by Ndiaye et al. (2017).

3.1 Generalized Gap Safe Screening

A natural choice would be to set $\mathcal{S} = \Delta_\mathbf{A}$ in Theorem 5 as it contains all possible feasible dual points, including the dual solution $\boldsymbol{\theta}^\star$. This choice is fine when D_λ is strongly concave on $\Delta_\mathbf{A}$, and when a strong concavity bound on this set can be derived. However, it may be necessary to further restrict $\Delta_\mathbf{A}$ in order to get the local strong concavity property, or to use a simpler shape for $\mathcal{S}$ in order to derive a strong concavity bound in closed-form. Hence, without loss of generality, we consider hereafter the set $\mathcal{S} = \Delta_\mathbf{A} \cap \mathcal{S}_0$, where $\mathcal{S}_0 \subseteq \mathbb{R}^m$ is such that $\boldsymbol{\theta}^\star \in \mathcal{S}_0$. A careful choice of $\mathcal{S}_0$ can turn out to be crucial in order to obtain the required local strong concavity property. For instance, this is the case for the $\beta = 1.5$ and Kullback-Leibler divergences, as discussed in Section 4.2.

Given a local strong concavity bound $\alpha_{\Delta_\mathbf{A} \cap \mathcal{S}_0}$ of D_λ over $\Delta_\mathbf{A} \cap \mathcal{S}_0$ (see Section 4.1.6), we revisit the Gap Safe dynamic screening approach (Algorithm 1) in the way it is presented in Algorithm 2. A notable difference is that the dual update (line 8) requires to output a point $\boldsymbol{\theta}$ in $\Delta_\mathbf{A} \cap \mathcal{S}_0$, i.e., $\boldsymbol{\Theta} : \mathbb{R}^n \to \Delta_\mathbf{A} \cap \mathcal{S}_0$ (instead of just $\Delta_\mathbf{A}$ in Algorithm 1). The reason is that, from Theorem 5, the ball $\mathcal{B}(\boldsymbol{\theta}, r)$ with r given at line 9 is ensured to be safe only if the center $\boldsymbol{\theta}$ belongs to the set on which the strong concavity bound has been computed, that is $\Delta_\mathbf{A} \cap \mathcal{S}_0$. In contrast, the intersection of $\mathcal{B}(\boldsymbol{\theta}, r)$ with $\mathcal{S}_0$ in the screening test (line 10) is not mandatory as $\mathcal{B}(\boldsymbol{\theta}, r)$ is itself a safe region. However, taking the intersection may lead to an even smaller set and, consequently, increase the number of screened variables.

Algorithm 2 Generalized Dynamic Gap Safe Screening (G-DGS):
$\hat{\mathbf{x}} = \texttt{GapSolver}(\mathbf{A}, \lambda, \mathcal{S}_0, \varepsilon_{\text{gap}})$

1: **Initialize** $\mathcal{A} = \mathcal{G}$, $\mathbf{x} \in \mathcal{C}$
2: **Set** $\boldsymbol{\eta}$ according to the solver
3: **Compute** $\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$ a strong concavity bound of D_λ on $\Delta_{\mathbf{A}} \cap \mathcal{S}_0$
4: **repeat**
5: — *Solver update restricted to preserved set* —
6: $\{\mathbf{x}_{\mathcal{A}}, \boldsymbol{\eta}\} \leftarrow \texttt{PrimalUpdate}(\mathbf{x}_{\mathcal{A}}, \mathbf{A}_{\mathcal{A}}, \lambda, \boldsymbol{\eta})$
7: — *Dynamic Screening* —
8: $\boldsymbol{\theta} \leftarrow \boldsymbol{\Theta}(\mathbf{x}) \in \Delta_{\mathbf{A}} \cap \mathcal{S}_0$. ▷ *Dual update*
9: $r \leftarrow \sqrt{\frac{2\,\text{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}}}$ ▷ *Safe radius*
10: $\mathcal{A} \leftarrow \{g \in \mathcal{A} \mid \max_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0} \overline{\Omega}_g(\phi(\mathbf{A}_g^\intercal \boldsymbol{\theta})) \geq 1\}$ ▷ *Screening test*
11: $\mathbf{x}_{\mathcal{A}^c} \leftarrow \mathbf{0}$
12: **until** $\text{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta}) < \varepsilon_{\text{gap}}$

3.2 Gap Safe Screening with Sphere Refinement

We now go one step further by observing that a Gap Safe sphere is a valid subset to invoke Theorem 5.

Corollary 6. *Let $\mathcal{S}_0$ be a safe region and $\mathcal{B}(\boldsymbol{\theta}, r)$ be a Gap Safe sphere for the primal-dual pair $(\mathbf{x}, \boldsymbol{\theta}) \in (\text{dom}(P_\lambda) \cap \mathcal{C}) \times (\Delta_{\mathbf{A}} \cap \mathcal{S}_0)$. If D_λ is $\alpha_{\mathcal{B}}$-strongly concave on $\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0$, then $\mathcal{B}(\boldsymbol{\theta}, \sqrt{2\,\text{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})/\alpha_{\mathcal{B}}})$ is a Gap Safe sphere.*

Proof Application of Theorem 5 with $\mathcal{S} = \mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0$. ■

Corollary 6 suggests that the radius r computed at line 9 of Algorithm 2 may be further reduced by computing a new strong concavity bound over $\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0$. We thus propose to replace the radius update at line 9 by an iterative refinement procedure, as depicted in Algorithm 3 (lines 14–18). The convergence of this refinement loop is guaranteed as stated in Proposition 7.

Proposition 7. *The sequence of safe radius generated by the refinement loop (lines 14–18) in Algorithm 3 converges.*

Proof Let $\mathcal{B}(\boldsymbol{\theta}^{\text{old}}, r^{\text{old}})$ be the current safe sphere at the end of a given iteration of the main loop. Let $(\mathbf{x}, \boldsymbol{\theta}) \in (\text{dom}(P_\lambda) \cap \mathcal{C}) \times (\Delta_{\mathbf{A}} \cap \mathcal{S}_0)$ be the updated primal-dual pair at the next iteration. Clearly, we have $\mathcal{B}(\boldsymbol{\theta}^{\text{old}}, r^{\text{old}}) \subseteq \mathcal{B}(\boldsymbol{\theta}^{\text{old}}, r)$ with $r = \max(r^{\text{old}}, \|\boldsymbol{\theta} - \boldsymbol{\theta}^{\text{old}}\|)$. Hence, $\mathcal{B}(\boldsymbol{\theta}^{\text{old}}, r)$ is also a safe sphere and, by construction, $\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}^{\text{old}}, r)$. Then, it follows from Theorem 5 (with $\mathcal{S} = \mathcal{B}(\boldsymbol{\theta}^{\text{old}}, r) \cap \mathcal{S}_0$) that $\mathcal{B}(\boldsymbol{\theta}, r')$ with $r' = \sqrt{2\,\text{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})/\alpha_{\mathcal{B}(\boldsymbol{\theta}^{\text{old}}, r) \cap \mathcal{S}_0}}$ is a safe sphere. This shows that, at line 15 of Algorithm 3, the computed radius r is such that $\mathcal{B}(\boldsymbol{\theta}, r)$ is a safe sphere. Then Corollary 6 combined with the "min" at line 17 ensures that the refinement loop builds a sequence of nested Gap Safe spheres (i.e., with decreasing radius), all centred in $\boldsymbol{\theta}$. As the radius is bounded below by 0, the proof is completed. ■

Algorithm 3 Refined Dynamic Gap Safe Screening (R-DGS) :
$\hat{\mathbf{x}} = \texttt{GAPSolver}(\mathbf{A}, \lambda, \mathcal{S}_0, \varepsilon_{\text{gap}}, \varepsilon_r)$

1: **Initialize** $\mathcal{A} = \mathcal{G}$, $\mathbf{x} \in \mathcal{C}$
2: **Set** $\boldsymbol{\eta}$ according to the solver
3: **Compute** $\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$ a strong concavity bound of D_λ on $\Delta_{\mathbf{A}} \cap \mathcal{S}_0$
4: — *Construction of an initial safe sphere* $\mathcal{B}(\boldsymbol{\theta}, r)$ —
5: $\boldsymbol{\theta} \leftarrow \boldsymbol{\Theta}(\mathbf{x}) \in (\Delta_{\mathbf{A}} \cap \mathcal{S}_0)$
6: $r \leftarrow \sqrt{\frac{2\,\text{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}}}$
7: — *Main loop* —
8: **repeat**
9: — *Solver update restricted to preserved set* —
10: $\{\mathbf{x}_{\mathcal{A}}, \boldsymbol{\eta}\} \leftarrow \texttt{PrimalUpdate}(\mathbf{x}_{\mathcal{A}}, \mathbf{A}_{\mathcal{A}}, \lambda, \boldsymbol{\eta})$
11: — *Dynamic Screening (adaptive local variant)* —
12: $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}$
13: $\boldsymbol{\theta} \leftarrow \boldsymbol{\Theta}(\mathbf{x}) \in (\Delta_{\mathbf{A}} \cap \mathcal{S}_0)$ ▷ *Dual update*
14: $r \leftarrow \max(r, \|\boldsymbol{\theta} - \boldsymbol{\theta}^{\text{old}}\|)$ ▷ *Initialize safe radius*
15: $r \leftarrow \sqrt{\frac{2\,\text{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha_{\mathcal{B}(\boldsymbol{\theta}^{\text{old}}, r) \cap \mathcal{S}_0}}}$
16: **repeat** ▷ *Refine safe radius*
17: $r \leftarrow \min\left(r, \sqrt{\frac{2\,\text{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha_{\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0}}}\right)$
18: **until** $\Delta r < \varepsilon_r$
19: $\mathcal{A} \leftarrow \{g \in \mathcal{A} \mid \max_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0} \overline{\Omega}_g(\phi(\mathbf{A}_g^\top \boldsymbol{\theta})) \geq 1\}$ ▷ *Screening test*
20: $\mathbf{x}_{\mathcal{A}^c} \leftarrow \mathbf{0}$
21: **until** $\text{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta}) < \varepsilon_{\text{gap}}$

The proposed radius refinement procedure can be interpreted as a two-step process. First, a Gap Safe sphere centred in $\boldsymbol{\theta}$ is computed from the previous one centred in $\boldsymbol{\theta}^{\text{old}}$ (lines 14–15). This step is required as Theorem 5 cannot be directly applied since the dual update (line 13) does not necessarily ensures $\boldsymbol{\theta}$ to belong to $\mathcal{B}(\boldsymbol{\theta}^{\text{old}}, r^{\text{old}})$. Then, the radius of this new Gap Safe sphere centred in $\boldsymbol{\theta}$ is iteratively reduced as long as the strong concavity bound improves (i.e., increases) when computed on the successively generated Gap Safe spheres (lines 16–18).

The computational overhead of this refinement procedure depends essentially on how efficiently the local strong concavity constant $\alpha_{\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0}$ can be computed which, in turn, depends on the dual function considered and $\mathcal{S}_0$. We shall show in Section 4 that this task can be done efficiently (in constant time) for several objective functions of practical interest.

4. Notable Particular Cases

In this section, we apply the proposed generic framework to ℓ_1-regularized problems[2] with some pertinent data-fidelity functions, namely: the quadratic distance, β-divergences with $\beta = 1.5$ and $\beta = 1$ (Kullback-Leibler divergence) and the Logistic regression objective. The β-divergence (Basu et al., 1998) is a family of cost functions parametrized by the a scalar β which is largely used in the context of Non-negative Matrix Factorization (NMF) with prominent application in audio (Févotte et al., 2018) and hyperspectral image processing (Févotte and Dobigeon, 2015). It covers as special cases the quadratic distance ($\beta = 2$) and the Kullback-Leibler divergence ($\beta = 1$). These examples are particularly interesting as they encompass the three following scenarios.

1. The dual cost function is only locally (*not* globally) strongly concave. The standard Gap Safe screening approach is not applicable while the proposed extension is. This is the case for β-divergences with $\beta \in [1, 2)$.

2. The dual cost function is globally strongly concave, but improved local strong-concavity bounds can be derived. We thus expect the proposed approach to improve over the standard Gap Safe screening. This is the case for the logistic regression.

3. The dual cost function is globally strongly concave and the global constant α cannot be improved locally. Here the proposed approach reduces to the standard Gap Safe screening. This is the case for the quadratic distance.

To highlight the specificities of each of these problems (e.g., the set $\mathcal{C}$, assumptions on $\mathbf{A}$ and the input data $\mathbf{y}$) we formalize them below:

- Quadratic distance: For $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\lambda > 0$,

$$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{m} (y_i - [\mathbf{A}\mathbf{x}]_i)^2 + \lambda \|\mathbf{x}\|_1. \tag{15}$$

- β-divergence with $\beta = 1.5$ (hereafter denoted $\beta_{1.5}$-divergence): For $\mathbf{y} \in \mathbb{R}_+^m$, $\mathbf{A} \in \mathbb{R}_+^{m \times n}$, $\lambda > 0$, and $\epsilon > 0$,

$$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}_+^n} \frac{4}{3} \sum_{i=1}^{m} \left(y_i^{3/2} + \frac{1}{2}([\mathbf{A}\mathbf{x}]_i + \epsilon)^{3/2} - \frac{3}{2} y_i ([\mathbf{A}\mathbf{x}]_i + \epsilon)^{1/2} \right) + \lambda \|\mathbf{x}\|_1 \tag{16}$$

- Kullback-Leibler divergence: For $\mathbf{y} \in \mathbb{R}_+^m$, $\mathbf{A} \in \mathbb{R}_+^{m \times n}$, $\lambda > 0$, and $\epsilon > 0$,

$$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}_+^n} \sum_{i=1}^{m} \left(y_i \log \left(\frac{y_i}{[\mathbf{A}\mathbf{x}]_i + \epsilon} \right) + [\mathbf{A}\mathbf{x}]_i + \epsilon - y_i \right) + \lambda \|\mathbf{x}\|_1 \tag{17}$$

2. Since the focus of this paper is more on the constraint set and data-fidelity term, we take a simple ℓ_1-norm as our default regularizer in all the discussed examples. Other regularizations have been explored by Ndiaye et al. (2017) and their results can be easily combined with the data-fidelity terms explored here.

- Logistic regression: For $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, and $\lambda > 0$,

$$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} \sum_{i=1}^{m} \left(\log\left(1 + \mathrm{e}^{[\mathbf{A}\mathbf{x}]_i}\right) - y_i [\mathbf{A}\mathbf{x}]_i \right) + \lambda \|\mathbf{x}\|_1 \tag{18}$$

We also assume in all cases that $\mathbf{A}$ has no all-zeros row, which is a natural assumption since otherwise the i-th entry y_i becomes irrelevant to the optimisation problem and can simply be removed (along with the corresponding row in $\mathbf{A}$).

All the quantities required to deploy Algorithm 2 and 3 on these problems are reported in Table 1. Although calculation details are deferred to Appendix D, we supply the main (generic) ingredients of these derivations in Section 4.1.

4.1 Useful Quantities

4.1.1 Regularization and dual norm

Let us instantiate some of the generic expressions in Section 2 for the case considered in this paper: $\Omega(\mathbf{x}) = \|\mathbf{x}\|_1$. Here, each group g is defined as an individual coordinate with $\mathcal{G} = \{\{1\}, \ldots, \{n\}\}$.

$$\begin{array}{ll} \Omega(\mathbf{x}) = \|\mathbf{x}\|_1 & \overline{\Omega}(\mathbf{x}) = \|\mathbf{x}\|_\infty = \max_{j \in [n]} |x_j| \\ \Omega_g(\mathbf{x}_g) = |x_g| & \overline{\Omega}_g(\mathbf{x}_g) = |x_g| \end{array} \tag{19}$$

4.1.2 Maximum Regularization Parameter

We call *maximum regularization parameter* the parameter $\lambda_{\max} > 0$ such that for all $\lambda \geq \lambda_{\max}$ the zero vector is a solution of (1). The following result generalizes (Ndiaye et al., 2017, Proposition 4) to the framework considered in this paper (see Appendix C.1 for a proof).

Proposition 8. $\lambda_{\max} = \overline{\Omega}\left(\phi\left(-\mathbf{A}^\mathsf{T} \nabla F(\mathbf{0})\right)\right)$.

ℓ_1-norm case. For $\Omega(\mathbf{x}) = \|\mathbf{x}\|_1$, we have

$$\lambda_{\max} := \|\phi\left(-\mathbf{A}^\mathsf{T} \nabla F(\mathbf{0})\right)\|_\infty = \begin{cases} \|\mathbf{A}^\mathsf{T} \nabla F(\mathbf{0})\|_\infty & \text{if } \mathcal{C} = \mathbb{R}^n \\ \max\left(-\mathbf{A}^\mathsf{T} \nabla F(\mathbf{0})\right) & \text{if } \mathcal{C} = \mathbb{R}^n_+ \end{cases} \tag{20}$$

The instantiation of this expression for each case under consideration is provided in Table 1.

4.1.3 Dual Feasible Set

As in our examples $\Omega = \|\cdot\|_1$, we get from Theorem 1 and (19) that

$$\Delta_{\mathbf{A}} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid |\phi([\mathbf{A}^\mathsf{T}\boldsymbol{\theta}]_j)| \leq 1, \ \forall j \in [n]\} \cap \operatorname{dom}(D_\lambda) \tag{21}$$

$$= \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \|\phi(\mathbf{A}^\mathsf{T}\boldsymbol{\theta})\|_\infty \leq 1\} \cap \operatorname{dom}(D_\lambda), \tag{22}$$

where we recall that $\phi(x) = x$ for problems where $\mathcal{C} = \mathbb{R}^n$ and $\phi(x) = [x]^+$ when $\mathcal{C} = \mathbb{R}^n_+$. Note that, because $|[z]^+| = [z]^+$ we have $\|\phi(\mathbf{A}^\mathsf{T}\boldsymbol{\theta})\|_\infty = \max_{j \in [n]}([\mathbf{A}^\mathsf{T}\boldsymbol{\theta}]_j)$ when $\mathcal{C} = \mathbb{R}^n_+$. Finally, the definition of $\operatorname{dom}(D_\lambda)$ is provided in Table 1 for each considered problem and details can be found in Appendix D.

4.1.4 Dual Update

At line 8 (resp., line 13) of Algorithm 2 (resp., Algorithm 3), one needs to compute a dual point $\boldsymbol{\Theta}(\mathbf{x}) \in (\Delta_{\mathbf{A}} \cap \mathcal{S}_0)$ from the current primal estimate $\mathbf{x}$. This point will define the center of the computed Gap Safe sphere. As discussed in Section 2.4, the function $\boldsymbol{\Theta} : \mathbb{R}^n \to \Delta_{\mathbf{A}} \cap \mathcal{S}_0$ should be such that

1. it can be evaluated efficiently (to limit the computational overhead);
2. $\|\boldsymbol{\Theta}(\mathbf{x}) - \boldsymbol{\theta}^\star\|$ is as small as possible (to reduce the sphere radius).

Given a primal point $\mathbf{x}$, the standard practice to compute a dual feasible point is to rescale the quantity $\nabla F(\mathbf{A}\mathbf{x})$ (Bonnefoy et al., 2015; Ndiaye et al., 2017). The rationale behind this choice is that the dual solution is a scaled version of $\nabla F(\mathbf{A}\mathbf{x}^\star)$ (see optimality condition (7)). We formalize this scaling procedure in Lemma 9 whose proof is given in Appendix C.2.

Lemma 9. *Assume that* $\mathrm{dom}(D_\lambda)$ *is stable by contraction and let* $\boldsymbol{\Xi} : \mathbb{R}^m \to \mathbb{R}^m$ *be the scaling operator defined by*

$$\boldsymbol{\Xi}(\mathbf{z}) := \frac{\mathbf{z}}{\max\left(\overline{\Omega}(\phi(\mathbf{A}^\mathsf{T}\mathbf{z})), 1\right)}. \tag{23}$$

Then, for any point $\mathbf{z} \in \mathrm{dom}(D_\lambda)$*, we have* $\boldsymbol{\Xi}(\mathbf{z}) \in \Delta_{\mathbf{A}}$*. Moreover, for any primal point* $\mathbf{x} \in \mathrm{dom}(P_\lambda)$*, we have that* $\mathbf{z} = (-\nabla F(\mathbf{A}\mathbf{x})/\lambda) \in \mathrm{dom}(D_\lambda)$ *and therefore*

$$\boldsymbol{\Xi}(-\nabla F(\mathbf{A}\mathbf{x})/\lambda) \in \Delta_{\mathbf{A}}. \tag{24}$$

Finally, if $F \in C^1$*, then* $\boldsymbol{\Xi}(-\nabla F(\mathbf{A}\mathbf{x})/\lambda) \to \boldsymbol{\theta}^\star$ *as* $\mathbf{x} \to \mathbf{x}^\star$*.*

From Lemma 9, we obtain a simple and cheap scaling procedure (the gradient $\nabla F(\mathbf{A}\mathbf{x})$ being often computed by the primal solver) that allows to obtain a dual feasible point $\boldsymbol{\Xi}(\mathbf{x}) \in \Delta_{\mathbf{A}}$ from any primal point $\mathbf{x} \in \mathrm{dom}(P_\lambda)$. Hence, when $\mathcal{S}_0 = \mathbb{R}^m$ one can simply set $\boldsymbol{\Theta}(\mathbf{x}) = \boldsymbol{\Xi}(-\nabla F(\mathbf{A}\mathbf{x})/\lambda)$. For other choices of $\mathcal{S}_0$, $\boldsymbol{\Xi}$ can be a starting point to define $\boldsymbol{\Theta}$ (see Table 1).

ℓ_1-norm case. For $\Omega(\mathbf{x}) = \|\mathbf{x}\|_1$, we have

$$\boldsymbol{\Xi}(\mathbf{z}) := \frac{\mathbf{z}}{\max\left(\|\phi(\mathbf{A}^\mathsf{T}\mathbf{z})\|_\infty, 1\right)}. \tag{25}$$

4.1.5 Sphere Test

The safe screening rule in Proposition 3 takes the following form when the safe region is a ℓ_2-ball, $\mathcal{B}(\boldsymbol{\theta}, r)$, with center $\boldsymbol{\theta}$ and radius r:

$$\begin{aligned}
\max_{\boldsymbol{\theta}' \in \mathcal{B}(\boldsymbol{\theta}, r)} \overline{\Omega}_g(\phi(\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta}')) &= \max_{\mathbf{u} \in \mathcal{B}(\mathbf{0},1)} \overline{\Omega}_g\left(\phi\left(\mathbf{A}_g^\mathsf{T}(\boldsymbol{\theta} + r\mathbf{u})\right)\right) \\
&\leq \max_{\mathbf{u} \in \mathcal{B}(\mathbf{0},1)} \overline{\Omega}_g\left(\phi(\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta}) + r\phi(\mathbf{A}_g^\mathsf{T}\mathbf{u})\right) \\
&\leq \overline{\Omega}_g\left(\phi(\mathbf{A}_g^\mathsf{T}\boldsymbol{\theta})\right) + r \max_{\mathbf{u} \in \mathcal{B}(\mathbf{0},1)} \overline{\Omega}_g\left(\phi(\mathbf{A}_g^\mathsf{T}\mathbf{u})\right)
\end{aligned} \tag{26}$$

where we used the subadditivity and homogeneity of the operator ϕ and the norm $\overline{\Omega}_g$. The resulting screening rule for the g-th group reads:

$$\overline{\Omega}_g\left(\phi(\mathbf{A}_g^\top \boldsymbol{\theta})\right) + r \max_{\mathbf{u}\in\mathcal{B}(\mathbf{0},1)} \overline{\Omega}_g\left(\phi(\mathbf{A}_g^\top \mathbf{u})\right) < 1 \implies \mathbf{x}_g^\star = \mathbf{0}. \tag{27}$$

ℓ_1-norm case. For $\Omega(\mathbf{x}) = \|\mathbf{x}\|_1$, the screening rule in (27) simplifies as:

$$\left|\phi(\mathbf{a}_j^\top \boldsymbol{\theta})\right| + r\|\mathbf{a}_j\| < 1 \implies x_j^\star = 0. \tag{28}$$

Finally, let us emphasise that in Algorithms 2 and 3, the safe region is $\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0$. Although the test (28) remains valid, a specific test that accounts for $\mathcal{S}_0$ (when $\mathcal{S}_0 \neq \mathbb{R}^m$) can increase the number of screened variables (see for instance Proposition 37 in Appendix D.4.5 for the KL-divergence).

4.1.6 Strong Concavity Bounds

The strong concavity parameters can be determined by upper-bounding the eigenvalues of the Hessian matrix $\nabla^2 D_\lambda(\boldsymbol{\theta})$. Indeed, a twice-differentiable function is strongly concave with constant $\alpha > 0$ if and only if its Hessian's eigenvalues are majorized by $-\alpha$, i.e., all eigenvalues are strictly negative with modulus greater or equal to α (Hiriart-Urruty and Lemaréchal, 1993a, Chapter IV, Theorem 4.3.1). In all considered particular cases, the resulting dual function is twice differentiable (as detailed in Appendix D). In Proposition 10, we derive the local strong concavity of D_λ on a convex set $\mathcal{S} \in \mathbb{R}^m$.

Proposition 10. *Assume that D_λ given in Theorem 1 is twice differentiable. Let $\mathcal{S} \in \mathbb{R}^m$ be a convex set and $\mathcal{I} = \{i \in [m] : \forall(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \mathcal{S}^2, \theta_i = \theta_i'\}$ a (potentially empty) set of coordinates in which $\mathcal{S}$ reduces to a singleton. Then, D_λ is $\alpha_\mathcal{S}$-strongly concave on $\mathcal{S}$ if and only if*

$$0 < \alpha_\mathcal{S} \leq \min_{i\in\mathcal{I}^\complement} -\sup_{\boldsymbol{\theta}\in\mathcal{S}} \sigma_i(\theta_i), \tag{29}$$

where $\sigma_i(\theta_i) = -\lambda^2 (f_i^*)''(\lambda\theta_i)$ is the (negative) i-th eigenvalue of the Hessian matrix $\nabla^2 D_\lambda(\boldsymbol{\theta})$.

Proof See Appendix C.3. ■

From Proposition 10, we define a general recipe to derive a strong concavity bound $\alpha_\mathcal{S}$ of D_λ on $\mathcal{S} \subset \mathbb{R}^m$:

1. Compute the eigenvalues $\sigma_i(\theta_i)$ of the Hessian of the dual function $\nabla^2 D_\lambda(\boldsymbol{\theta})$.
2. Upper-bound the i-th eigenvalue $\sigma_i(\theta_i)$ over the set $\mathcal{S}$.
3. Take the minimum over $i \in \mathcal{I}^\complement$ of the opposite (negative) upper-bounds.

The bound $\alpha_\mathcal{S}$ obtained via the above procedure is valid if it is strictly positive. For each considered loss, we provide in Table 1 strong concavity bounds for three different choices for the set $\mathcal{S}$: 1) $\mathcal{S} = \mathbb{R}^m$ (global bound), 2) $\mathcal{S} = \Delta_\mathbf{A} \cap \mathcal{S}_0$ and 3) $\mathcal{S} = \mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0$ (local bounds). The choice of the set $\mathcal{S}_0$ is discussed in Section 4.2 and details concerning the derivation of these bounds are provided in Appendix D.

4.2 Discussion

The quadratic case is quite straightforward, as one can easily verify that the dual function is globally strongly-concave with constant $\alpha = \lambda^2$ and we can simply take $\mathcal{S}_0 = \mathbb{R}^m$ (see Appendix D.1). Moreover, because all eigenvalues of the Hessian $\nabla^2 D_\lambda(\boldsymbol{\theta})$ are constant and equal to $-\lambda^2$ for all $\boldsymbol{\theta}$, every local strong-concavity bound coincides with the global one.

In the following sections, we discuss the particularities of the three remaining cases: $\beta_{1.5}$-divergence, Kullback-Leibler divergence, and logistic regression.

The following set will be useful to define $\mathcal{S}_0$ for both $\beta_{1.5}$ and KL divergences.

Definition 11. *We denote $\mathcal{I}_0 \in [m]$ the set of coordinates for which the input data $\mathbf{y}$ equals zero:*

$$\mathcal{I}_0 := \{i \in [m] \mid y_i = 0\}$$

4.2.1 $\beta_{1.5}$ Divergence Case

In this case, the i-th eigenvalue of the Hessian $\nabla^2 D_\lambda(\boldsymbol{\theta})$ is given by (see Appendix D.3):

$$\sigma_i(\theta_i) = -\lambda^2 \left(\frac{(\lambda\theta_i)^2 + 2y_i}{\sqrt{(\lambda\theta_i)^2 + 4y_i}} - \lambda\theta_i \right). \tag{30}$$

One can see that the eigenvalues are all non-positive, but may vanish in two cases: 1) when $\theta_i \to +\infty$, 2) when $y_i = 0$ and $\theta_i \geq 0$.

This means that the corresponding dual function is not globally strongly concave. Moreover, one can show that the dual function is still not strongly concave when restricted to the dual feasible set $\Delta_{\mathbf{A}}$ (see Figure 1a). Therefore, the application of the proposed local approach requires the definition of some additional constraint set $\mathcal{S}_0$. Below, we show that these two problems can be fixed by setting

$$\mathcal{S}_0 = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \lambda\boldsymbol{\theta} \leq \mathbf{b}\}, \tag{31}$$

where $\mathbf{b}$ is such that $\mathbf{b}_{\mathcal{I}_0} < 0$ and $\lambda\boldsymbol{\theta}^\star \leq \mathbf{b}$.

- The first problem is that the eigenvalues $\{\sigma_i\}_i$ are increasing functions of $\{\theta_i\}_i$ (see Proposition 25) and tend to zero as $\theta_i \to +\infty$. This can be prevented by restricting ourselves to a subset that upper bounds $\boldsymbol{\theta}$. Unfortunately, it is not sufficient to restrict $\boldsymbol{\theta}$ to the feasible set $\Delta_{\mathbf{A}} = \{\boldsymbol{\theta} \mid \mathbf{A}^\mathsf{T}\boldsymbol{\theta} \leq 1\}$ as illustrated in Figure 1a (one may indefinitely increase the value of a given coordinate by reducing the value of the others). However, the introduction of $\mathcal{S}_0$ defined in (31) fix this issue, as illustrated in Figure 1b.

- The second problem is that the eigenvalues $\{\sigma_i\}_i$ are equal to 0 when $i \in \mathcal{I}_0$ and $\theta_i \geq 0$. Again, one can easily see that the introduction of $\mathcal{S}_0$ defined in (31) fix this issue (thanks to $\mathbf{b}_{\mathcal{I}_0} < 0$).

The bound $\mathbf{b} = (\mathbf{y} - \epsilon)/\sqrt{\epsilon}$ is simple and valid choice. Indeed, we get from the primal-dual link in optimality condition (7) that $\lambda\theta_i^\star \leq (\mathbf{y} - \epsilon)/\sqrt{\epsilon}$ (see equation (103) in Appendix D.3). Moreover, we have, for all $i \in \mathcal{I}_0$, $b_i = (y_i - \epsilon)/\sqrt{\epsilon} = -\sqrt{\epsilon} < 0$. Yet, in Table 1 and Proposition 26, we provide the expression of a more complex but finer bound $\mathbf{b}$ that leads

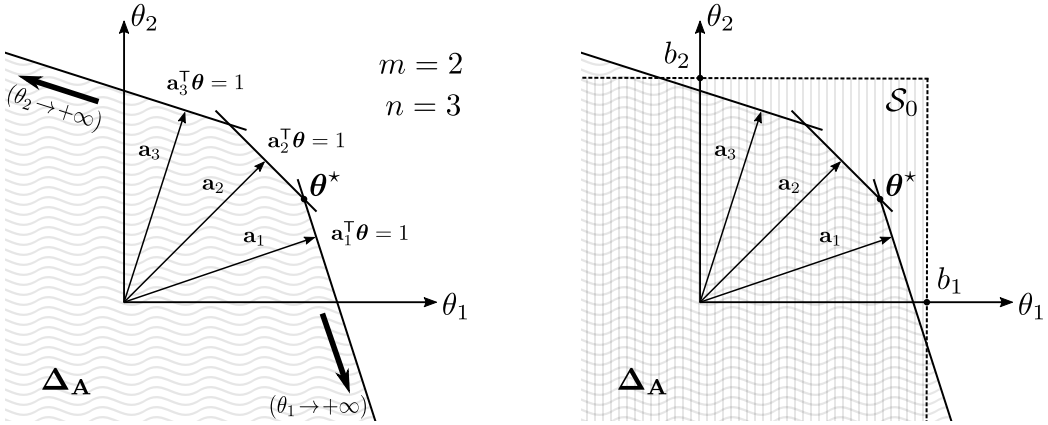


Figure 1: Motivation for the choice of $\mathcal{S}_0$ set in the $\beta_{1.5}$-divergence case with $\mathcal{I}_0 = \emptyset$. a) The thick arrows show directions along which $\theta_i \to +\infty$ and, as a consequence, the singular values of the Hessian $\nabla^2 D_\lambda(\boldsymbol{\theta})$ tend to zero. b) Upper-bounding θ_i solves the problem.

to improved strong concavity constants (which are particularly relevant for Algorithm 2). Finally, note that $\boldsymbol{\theta}^\star \in \mathcal{S}_0$, as required.

The resulting local strong concavity bounds $\alpha_{\Delta_\mathbf{A} \cap \mathcal{S}_0}$ (Proposition 27) and $\alpha_{\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0}$ (Proposition 28) are presented in Table 1. The dual update discussed in Section 4.1 is also adapted in view of $\mathcal{S}_0$ (Proposition 29).

4.2.2 $\beta = 1$ Kullback-Leibler Divergence Case

For the Kullback-Leibler divergence case, the i-th eigenvalue of the Hessian $\nabla^2 D_\lambda(\boldsymbol{\theta})$ is given by (see Appendix D.4) :

$$\sigma_i(\theta_i) = -\lambda^2 \frac{y_i}{(1 + \lambda\theta_i)^2}. \tag{32}$$

One can see that the eigenvalues are all non-positive, but may vanish in two cases: 1) when $\theta_i \to +\infty$. 2) when $y_i = 0$. This means that, like before, the dual function is not globally strongly concave. The first problem is prevented when we restrict ourselves to any bounded set, which is the case for both $\Delta_\mathbf{A}$ and $\mathcal{B}(\boldsymbol{\theta}, r)$. We recall that in this particular case $\Delta_\mathbf{A} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \mathbf{A}^\top\boldsymbol{\theta} \leq 1,\ \boldsymbol{\theta} \geq -1/\lambda\}$, where the second inequality corresponds to the domain of the dual function. However, the case $y_i = 0$ remains a problem and, once again, the application of the proposed local approach requires the definition of some additional constraint set $\mathcal{S}_0$.

Fortunately, for coordinates $i \in \mathcal{I}_0$, the dual solution is trivially determined. Indeed, the primal-dual link in optimality condition (7) gives that $\theta_i^\star = -1/\lambda$ (see equation (130) in Appendix D.4). This leads us to the following constraint set (Proposition 33):

$$\mathcal{S}_0 = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\theta}_{\mathcal{I}_0} = -\mathbf{1}/\lambda\}. \tag{33}$$

Note that $\mathcal{S}_0$ reduces to a singleton on coordinates $i \in \mathcal{I}_0$ and, from Proposition 10, the corresponding eigenvalues of $\nabla^2 D_\lambda$ can be neglected. Moreover, $\boldsymbol{\theta}^\star \in \mathcal{S}_0$, as required.

This set $\mathcal{S}_0$ allows us to compute the local bounds $\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$ (Proposition 34) and $\alpha_{\mathcal{B}(\boldsymbol{\theta},r) \cap \mathcal{S}_0}$ (Proposition 35). Additionally, the dual update is adapted considering $\mathcal{S}_0$ in Proposition 36 and an improved screening test is defined in Proposition 37. All these quantities are reported in Table 1.

In a previous paper (Dantas et al., 2021), we addressed the KL particular case and proposed a local bound $\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$ for a screening approach similar to Algorithm 2. However, the bound proposed here is far superior than the previous one, as will be demonstrated in the experimental part. Moreover, the iterative refinement approach in Algorithm 3 was not a part of Dantas et al. (2021) and no bound $\alpha_{\mathcal{B}(\boldsymbol{\theta},r) \cap \mathcal{S}_0}$ was provided.

4.2.3 Logistic Case

In this case, the i-th eigenvalue of the Hessian $\nabla^2 D_\lambda(\boldsymbol{\theta})$ is given by (see Appendix D.2):

$$\sigma_i\left(\theta_i\right) = -\frac{\lambda^2}{(y_i - \lambda\theta_i)(1 - y_i + \lambda\theta_i)}. \tag{34}$$

Similarly to the quadratic case, the dual function is globally strongly concave as the eigenvalues never vanish. For this reason, we can simply take $\mathcal{S}_0 = \mathbb{R}^m$. On the other hand, differently from the quadratic case, the eigenvalues are not constant and the proposed local approach may lead to better strong-concavity bounds.

Indeed, the local strong-concavity bound $\alpha_{\Delta_{\mathbf{A}}}$ (Proposition 21) does improve upon the global bound depending on the regularization parameter λ. More precisely, it improves upon the global constant $\alpha = 4\lambda^2$ (see Table 1) for all

$$\lambda < \frac{1}{2\|\mathbf{A}^\dagger\|_1}.$$

where $\mathbf{A}^\dagger$ denotes the right pseudo-inverse of $\mathbf{A}$. It is important to mention that the proposed $\alpha_{\Delta_{\mathbf{A}}}$ uses the supplementary assumption that $\mathbf{A}$ is full rank with $\operatorname{rank}(\mathbf{A}) = \min(n, m)$. If this hypothesis is not fullfiled, one can always initialize Algorithms 2 and 3 with the global bound $\alpha = 4\lambda^2$. When deploying Algorithm 3, this bound will be automatically refined during the iterations thanks to the local bound $\alpha_{\mathcal{B}(\boldsymbol{\theta},r)}$ (Proposition 22).

5. Experiments

In this section, our objective is to evaluate the proposed techniques in a diverse range of scenarios and provide essential insights for efficiently applying such techniques to other problems and settings. To that end, a wide range of experiments have been performed but only a representative subset of the results are presented in this section for the sake of conciseness and readability. The complete Matlab code is made available by the authors[3] for the sake of reproducibility. Different simulation scenarios (not reported here) are also available for the interested reader.

Solvers. For each of the three problems described in Section 4, namely logistic regression, KL divergence, and $\beta_{1.5}$-divergence, we deploy some popular solvers from the literature in

3. Code available at: `https://github.com/cassiofragadantas/`

	$\beta = 2$ (quadratic)	$\beta = 1.5$	$\beta = 1$ (Kullback-Leibler)	Logistic
$F(\mathbf{Ax})$	$\frac{1}{2}\|\mathbf{y} - \mathbf{Ax}\|_2^2$	$\frac{4}{3}\|\mathbf{y}\|_{1.5}^{1.5} + \frac{2}{3}\|\mathbf{Ax} + \epsilon\|_{1.5}^{1.5}$ $-2\mathbf{y}^\mathsf{T}(\mathbf{Ax} + \epsilon)^{0.5}$	$\mathbf{y}^\mathsf{T}\log(\frac{\mathbf{y}}{\mathbf{Ax}+\epsilon}) + \mathbf{1}^\mathsf{T}(\mathbf{Ax} + \epsilon - \mathbf{y})$	$\mathbf{1}^\mathsf{T}\log\left(1 + \mathrm{e}^{\mathbf{Ax}}\right) - \mathbf{y}^\mathsf{T}\mathbf{Ax}$
$\mathcal{C}$	$\mathbb{R}^n$	$\mathbb{R}^n_+$	$\mathbb{R}^n_+$	$\mathbb{R}^n$
$D_\lambda(\boldsymbol{\theta})$	$\frac{1}{2}\left(\|\mathbf{y}\|_2^2 - \|\mathbf{y} - \lambda\boldsymbol{\theta}\|_2^2\right)$	$\frac{1}{6}\|\lambda\boldsymbol{\theta}\|_3^3 - \frac{1}{6}\|(\lambda\boldsymbol{\theta})^2 + 4\mathbf{y}\|_{1.5}^{1.5}$ $+\lambda\boldsymbol{\theta}^\mathsf{T}\mathbf{y} + \frac{4}{3}\|\mathbf{y}\|_{1.5}^{1.5} - \epsilon\lambda\boldsymbol{\theta}^\mathsf{T}\mathbf{1}$	$\mathbf{y}^\mathsf{T}\log(1 + \lambda\boldsymbol{\theta}) - \lambda\epsilon\boldsymbol{\theta}^\mathsf{T}\mathbf{1}$	$(\mathbf{y}-\lambda\boldsymbol{\theta}-1)^\mathsf{T}\log(1-\mathbf{y}+\lambda\boldsymbol{\theta})$ $-(\mathbf{y}-\lambda\boldsymbol{\theta})^\mathsf{T}\log(\mathbf{y}-\lambda\boldsymbol{\theta})$
$\mathrm{dom}(D_\lambda)$	$\mathbb{R}^m$	$\mathbb{R}^m$	$\{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\theta} \geq -\mathbf{1}/\lambda\}$	$\{\boldsymbol{\theta} \in \mathbb{R}^m \mid \mathbf{y} - 1 \leq \lambda\boldsymbol{\theta} \leq \mathbf{y}\}$
$\Delta_\mathbf{A}$	$\{\boldsymbol{\theta}\in\mathbb{R}^m \mid \|\mathbf{A}^\mathsf{T}\boldsymbol{\theta}\|_\infty \leq 1\}$	$\{\boldsymbol{\theta} \in \mathbb{R}^m \mid \mathbf{A}^\mathsf{T}\boldsymbol{\theta} \leq \mathbf{1}\}$	$\{\boldsymbol{\theta} \in \mathbb{R}^m \mid \mathbf{A}^\mathsf{T}\boldsymbol{\theta} \leq \mathbf{1}, \boldsymbol{\theta} \geq -\mathbf{1}/\lambda\}$	$\{\boldsymbol{\theta} \in \mathbb{R}^m \mid \|\mathbf{A}^\mathsf{T}\boldsymbol{\theta}\|_\infty \leq 1,$ $\mathbf{y} - 1 \leq \lambda\boldsymbol{\theta} \leq \mathbf{y}\}$
$\mathcal{S}_0$	$\mathbb{R}^m$	$\{\boldsymbol{\theta} \in \mathbb{R}^m \mid \lambda\boldsymbol{\theta} \leq \min(\mathbf{b}, \frac{\mathbf{y}-\epsilon}{\sqrt{\epsilon}})\}$ $b_i = \lambda \min_j\left(\frac{1-c\|\mathbf{a}_j\|_1}{a_{ij}}\right) + \lambda c$ $c = -\frac{1}{\lambda}\sqrt[3]{\frac{4\|\mathbf{y}\|_{1.5}^{1.5} + 2(m-1)\epsilon^{1.5} + 3\epsilon}{1-3\epsilon}}$	$\{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\theta}_{\mathcal{I}_0} = -\mathbf{1}/\lambda\}$ with $\mathcal{I}_0 = \{i \in [m] \mid y_i = 0\}$	$\mathbb{R}^m$
Primal-dual link	$\lambda\boldsymbol{\theta}^\star = \mathbf{y} - \mathbf{Ax}^\star$	$\lambda\boldsymbol{\theta}^\star = \frac{\mathbf{y}}{\sqrt{\mathbf{Ax}^\star + \epsilon}} - \sqrt{\mathbf{Ax}^\star + \epsilon}$	$\lambda\boldsymbol{\theta}^\star = \frac{\mathbf{y}}{\mathbf{Ax}^\star} - \mathbf{1}$	$\lambda\boldsymbol{\theta}^\star = \mathbf{y} - \frac{\mathrm{e}^{\mathbf{Ax}^\star}}{1+\mathrm{e}^{\mathbf{Ax}^\star}}$
$\lambda_{\max}$	$\|\mathbf{A}^\mathsf{T}\mathbf{y}\|_\infty$	$\max\left(\mathbf{A}^\mathsf{T}(\mathbf{y} - \epsilon)\right)/\sqrt{\epsilon}$	$\max\left(\mathbf{A}^\mathsf{T}(\mathbf{y} - \epsilon)\right)/\epsilon$	$\|\mathbf{A}^\mathsf{T}(\mathbf{y} - \frac{1}{2})\|_\infty$
Dual update $\boldsymbol{\Theta}(\mathbf{x}) \in \Delta_\mathbf{A} \cap \mathcal{S}_0$	$\boldsymbol{\Xi}\left((\mathbf{y} - \mathbf{Ax})/\lambda\right)$	$\min\left(\left[\boldsymbol{\Xi}\left(\frac{\mathbf{y}-\mathbf{Ax}-\epsilon}{\lambda\sqrt{\mathbf{Ax}+\epsilon}}\right)\right]_i,\ \frac{b_i}{\lambda},\ \frac{y_i-\epsilon}{\lambda\sqrt{\epsilon}}\right)$	$\begin{cases} \left[\boldsymbol{\Xi}\left(\frac{1}{\lambda}\left(\frac{\mathbf{y}}{\mathbf{Ax}+\epsilon} - \mathbf{1}\right)\right)\right]_i & \text{if } i \in \mathcal{I}_0^\complement \\ -\frac{1}{\lambda} & \text{if } i \in \mathcal{I}_0 \end{cases}$	$\boldsymbol{\Xi}\left(\frac{1}{\lambda}\left(\mathbf{y} - \frac{\mathrm{e}^{\mathbf{Ax}}}{1+\mathrm{e}^{\mathbf{Ax}}}\right)\right)$
$\nabla^2 D_\lambda(\boldsymbol{\theta})$	$-\lambda^2\mathrm{Diag}\left(1, \ldots, 1\right)$	$-\lambda^2\mathrm{Diag}\left(\left[\frac{(\lambda\theta_i)^2 + 2y_i}{\sqrt{(\lambda\theta_i)^2 + 4y_i}} - \lambda\theta_i\right]_i\right)$	$-\lambda^2\mathrm{Diag}\left(\left[\frac{y_i}{(1+\lambda\theta_i)^2}\right]_i\right)$	$-\lambda^2\mathrm{Diag}\left(\left[\frac{4}{1-4(\lambda\theta_i - y_i + \frac{1}{2})^2}\right]_i\right)$
α (global)	λ^2	–	–	$4\lambda^2$
$\alpha_{\Delta_\mathbf{A} \cap \mathcal{S}_0}$	λ^2	$\min_{i\in[m]} -\sigma_i\left(\frac{1}{\lambda}\min\left(b_i, \frac{y_i-\epsilon}{\sqrt{\epsilon}}\right)\right)$ $\sigma_i(\frac{z_i}{\lambda}) = -\lambda^2\left(\frac{z_i^2 + 2y_i}{\sqrt{z_i^2 + 4y_i}} - z_i\right)$	$\lambda^2 \min_{i\in\mathcal{I}_0^\complement} \frac{y_i}{\left(\min_j\left(\frac{\lambda + \|\mathbf{a}_j\|_1}{a_{ij}}\right)\right)^2}$	$\frac{4\lambda^2}{1-4\left(\min(\lambda\|\mathbf{A}^\dagger\|_1, \frac{1}{2}) - \frac{1}{2}\right)^2}$
$\alpha_{\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0}$	λ^2	$\min_{i\in[m]} -\sigma_i(d_i/\lambda)$ $d_i = \min\left(\lambda(\theta_i + r),\ b_i,\ \frac{y_i-\epsilon}{\sqrt{\epsilon}}\right)$	$\lambda^2 \min_{i\in\mathcal{I}_0^\complement} \frac{y_i}{(1+\lambda(\theta_i + r))^2}$	$\frac{4\lambda^2}{1-4([\min_i(\lvert\lambda\theta_i - y_i + \frac{1}{2}\rvert) - \lambda r]^+)^2}$
Screening test ($\Omega(\mathbf{x}) = \|\mathbf{x}\|_1$)	$\lvert\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}\rvert + r\|\mathbf{a}_j\|_2 < 1$	$\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta} + r\|\mathbf{a}_j\|_2 < 1$	$\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta} + r\|[\mathbf{a}_j]_{\mathcal{I}_0}\|_2 < 1$	$\lvert\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}\rvert + r\|\mathbf{a}_j\|_2 < 1$

Table 1: Screening with three instances of β-divergence and logistic regression.

their standard form (i.e., without screening) as well as within the three mentioned screening approaches: the existing dynamic Gap Safe approach in Algorithm 1 (when applicable) and the proposed screening methods with local strong concavity bounds in Algorithms 2 and 3. Three different categories of solvers are used: coordinate descent (CoD) (Friedman et al., 2010; Hsieh and Dhillon, 2011; Yuan et al., 2010), multiplicative update (MU) yielding from majorization-minimization (Févotte and Idier, 2011) and proximal gradient algorithms (Harmany et al., 2012). The chosen solvers and the corresponding variations for each particular case are listed in Table 2.

Data Sets. Well-suited data sets have been chosen according to each considered problem (see Table 2). The Leukemia binary classification data set (Golub et al., 1999)[4] is used for the logistic regression case. For the KL case, the NIPS papers word count data set (Globerson et al., 2007)[5] is considered. For the β-divergence case, we use the Urban hyperspectral image data set (Jia and Qian, 2007)[6], since the β-divergence (especially with $\beta = 1.5$) was reported to be well-suited for this particular type of data (Févotte and Dobigeon, 2015). Additional details are given in the following respective sections.

Problem	Solvers	Variations	Data
Logistic	CoD	No screening, Alg. 1 (DGS), Alg. 2 (G-DGS), Alg. 3 (R-DGS)	Binary classification (Leukemia data set)
KL	MU, CoD, Prox. Grad.	No screening, Alg. 2 (G-DGS), Alg. 3 (R-DGS)	Count data (NIPS papers data set)
$\beta = 1.5$	MU	No screening, Alg. 2 (G-DGS), Alg. 3 (R-DGS)	Hyperspectral data (Urban image)

Table 2: Experimental scenarios

Evaluation. To compare the tested approaches, two main performance measures are used

1. Screening rate: how many (and how quickly) inactive coordinates are identified and eliminated by the compared screening strategies.
2. Execution time: the impact of screening in accelerating the solver's convergence.

Table 3 specifies the explored values of two parameters with decisive impact in the performance measures. We set the "smoothing" parameter of $\beta_{1.5}$ and KL divergence to $\epsilon = 10^{-6}$. Remaining problem parameters are fixed by the choice of the data set: problem dimensions (m, n) and data distribution (both the input vector $\mathbf{y}$ and matrix $\mathbf{A}$).

This section is organized as follows: the particular cases of logistic regression, $\beta_{1.5}$-divergence and Kullback-Leibler divergence are treated respectively in Sections 5.1 to 5.3. Other worth-mentioning properties of the proposed approaches are discussed in Section 5.4, notably the robustness of Algorithm 3 to the initialization of the strong concavity bound.

4. Data set available at LIBSVM: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`
5. Data set available at: `http://ai.stanford.edu/~gal/data.html`
6. Data set available at: `https://rslab.ut.ac.ir/data`

Parameter	Range
Regularization ($\lambda/\lambda_{\max}$)	$[10^{-3}, 1]$
Stopping criterion (ϵ_{gap})	$\{10^{-7}, 10^{-5}\}$

Table 3: Simulation parameters and explored values.

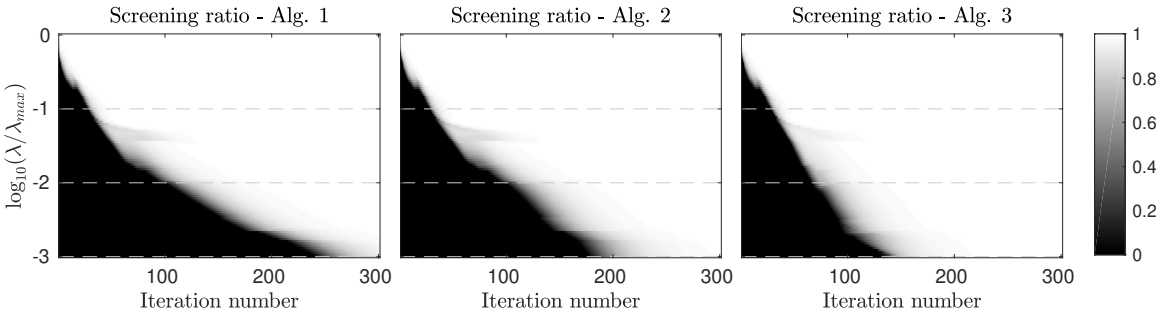
5.1 Logistic Regression

This example is particularly interesting as it allows to compare all screening approaches (Algorithms 1 to 3). The classic Leukemia binary classification data set is used in the experiments,[7] leading to a matrix $\mathbf{A}$ with dimensions $(m \times n) = (71 \times 7129)$, whose columns are re-normalized to unit-norm. Vector $\mathbf{y}$ contains the binary labels $\{0, 1\}$ of each sample. A coordinate descent algorithm (Yuan et al., 2010) is used to optimize problem (18)—same as used by Ndiaye et al. (2017).

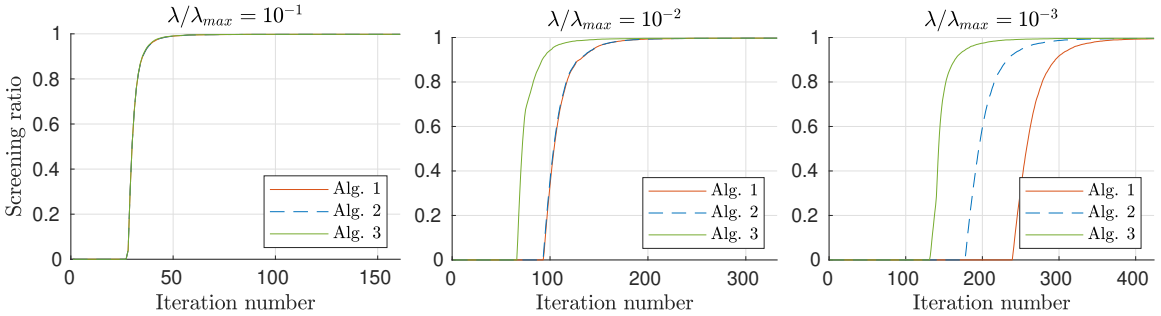
Figures 2a and 2b show the screening ratio (number of screened coordinates divided by the total number of coordinates) as a function of the iteration number, Figures 2c and 2d show the convergence rate (duality gap) as a function of the execution time and Figure 3 depicts relative execution times (where the solver without screening is taken as the reference) as a function of the regularization parameter. Figures 2a and 2b show a clear hierarchy between Algorithms 1, 2 and 3 in terms of screening performance, from worst to best. The difference is particularly pronounced at lower regularizations—cf. plot with $\lambda/\lambda_{\max} = 10^{-3}$. As regularization grows, Algorithms 1 and 2 become equivalent ($\lambda/\lambda_{\max} = 10^{-2}$) and later, around $\lambda/\lambda_{\max} = 10^{-1}$, all approaches become equivalent. The first mentioned transition point, where Algorithms 1 and 2 become equivalent, can be theoretically predicted at $\lambda = (2\|\mathbf{A}^\dagger\|_1)^{-1}$, as discussed in Section 4.2.3. In the reported example this corresponds to $\lambda = 1.2 \times 10^{-2}\lambda_{\max}$. This threshold is depicted as a vertical dotted line in Figure 3 and it accurately matches the experimental results.

The discussed screening performances translate quite directly in terms of execution times, as shown in Figures 2c and 2d. Figure 3 shows the relative execution times, where the basic solver (without screening) is taken as the reference, for a range of regularisation values at given convergence threshold ($\epsilon_{\text{gap}} = \{10^{-5}, 10^{-7}\}$). A typical behavior of screening techniques is observed: the smaller the convergence tolerance, the more advantageous screening becomes. Indeed, a smaller convergence tolerance leads to a larger number of iterations in the final optimization stage, where most coordinates are already screened out. A considerable speedup is obtained with Algorithm 3 in comparison to the other approaches in a wide range of regularization values and, in worst-case scenario, it is equivalent to other screening approaches (i.e., no significant overhead is observed). These results indicate that the proposed Algorithm 3 (R-DGS) should be preferred over the other approaches in this particular problem.

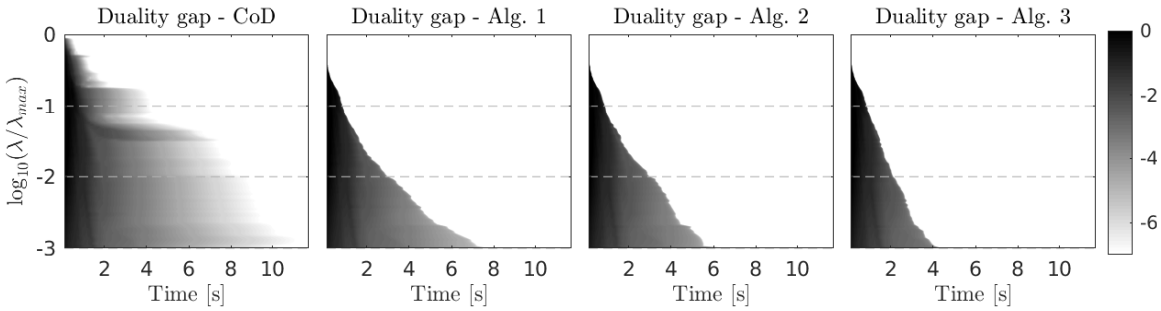
7. Sample number 17 was removed from the original Leukemia data set in order to improve the conditioning of matrix $\mathbf{A}$, which would be nearly singular otherwise and the proposed bound $\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$ in Proposition 27 would reduce to the global $\alpha_{\mathbb{R}^m}$, leading to uninteresting results (although still technically correct).



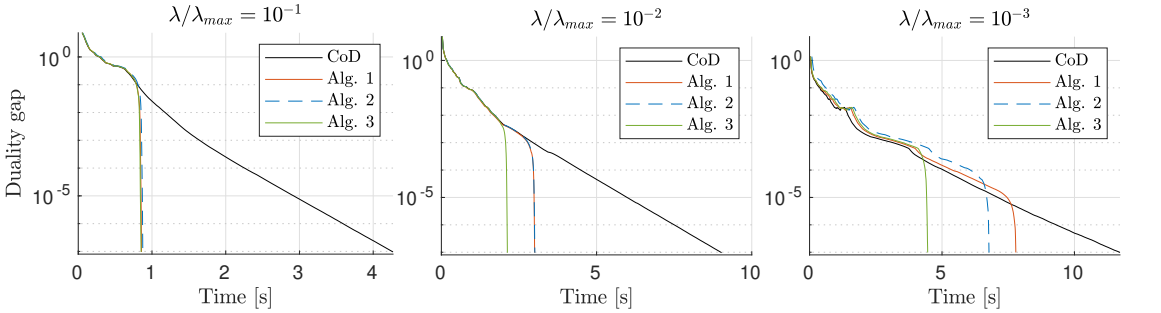
(a) Screening ratio against iterations for $\lambda/\lambda_{\max} \in [10^{-1}, 10^{-3}]$ (the lighter, the more screened coordinates), for Algorithms 1 to 3.



(b) Screening ratio against iterations for fixed regularization $\lambda/\lambda_{\max} = \{10^{-1}, 10^{-2}, 10^{-3}\}$, corresponding to the dotted slices in Figure 2a.



(c) Convergence rate (duality gap) against execution time for $\lambda/\lambda_{\max} \in [10^{-1}, 10^{-3}]$ (the lighter, the closer to convergence). Left to right: Coordinate Descent solver alone and in Algorithms 1 to 3.



(d) Duality gap against time for fixed regularization $\lambda/\lambda_{\max} = \{10^{-1}, 10^{-2}, 10^{-3}\}$, corresponding to dotted slices in Figure 2c.

Figure 2: Sparse Logistic regression of Leukemia data set using Coordinate Descent and screening.

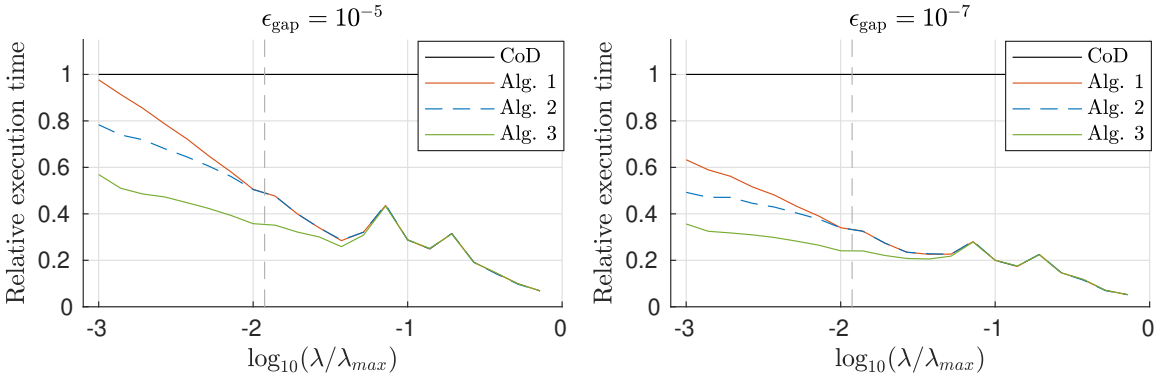


Figure 3: Sparse Logistic regression on Leukemia data set using Coordinate Descent and screening. Relative execution times for $\lambda/\lambda_{\max} \in [10^{-3}, 1]$ (the smaller value, the faster) with convergence criterion $\epsilon_{\text{gap}} = 10^{-5}$ (left) and $\epsilon_{\text{gap}} = 10^{-7}$ (right).

5.2 $\beta_{1.5}$ Divergence

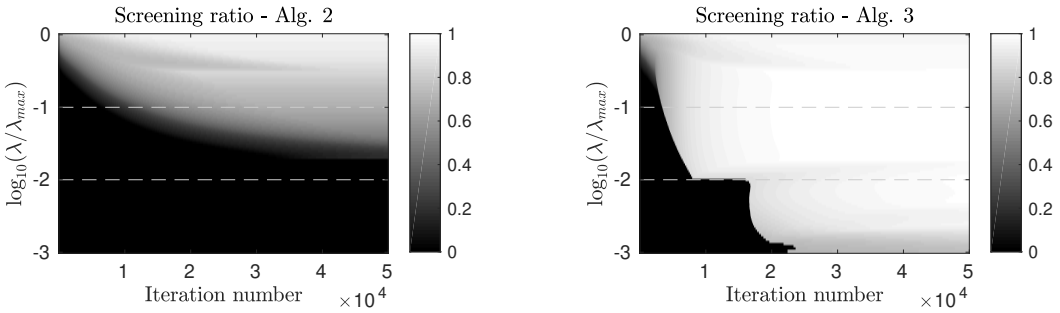
In this section, we use the Urban data set which is a (307×307)-pixel hyperspectral image with 162 spectral bands per pixel. For the experiments, the input signal $\mathbf{y} \in \mathbb{R}^{162}$ is given by a randomly-selected pixel from the image and the dictionary matrix $\mathbf{A} \in \mathbb{R}^{162\times 5000}$ is made of a uniformly-distributed random subset of 5000 of the remaining pixels. Therefore, the goal is to sparsely reconstruct a given pixel (in $\mathbf{y}$) as a sparse combination of other pixels from the same image, akin to archetypal analysis (Cutler and Breiman, 1994). The multiplicative MM algorithm described in (Févotte and Idier, 2011) is used to solve problem (16).

Figures 4a and 4b show the screening performance for Algorithms 2 and 3. Here, there is an even more pronounced difference between the two approaches (compared to the logistic regression case in Section 5.1) for the entire range of regularization values. Note that Algorithm 2 does not screen at all for $\lambda/\lambda_{\max} \leq 10^{-2}$, as opposed to Algorithm 3. This means that the refinement strategy in the latter approach manages to significantly improve, along the iterations, the initial strong-concavity bound (kept constant by the former). Even in highly-regularized scenarios, the screening ratio grows significantly faster with Algorithm 3.

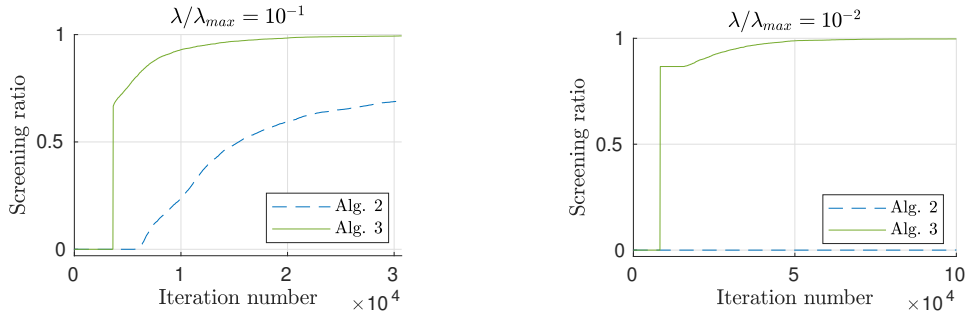
A similar behavior is observed in Figures 4c and 4d regarding execution times. Because no screening is performed by Algorithm 2 at regularization $\lambda/\lambda_{\max} = 10^{-2}$ (and below), no speedup is obtained w.r.t. the basic solver—there is even a slight overhead due to unfruitful screening tests calculations. Algorithm 2 only provides speedup over the basic solver for more regularized scenarios. Algorithm 3, in turn, provides acceleration over the entire regularization range and significantly outperforms both the basic solver and Algorithm 2. The previous observations are summarized in Figure 4e for normalized execution times.

5.3 Kullback-Leibler Divergence

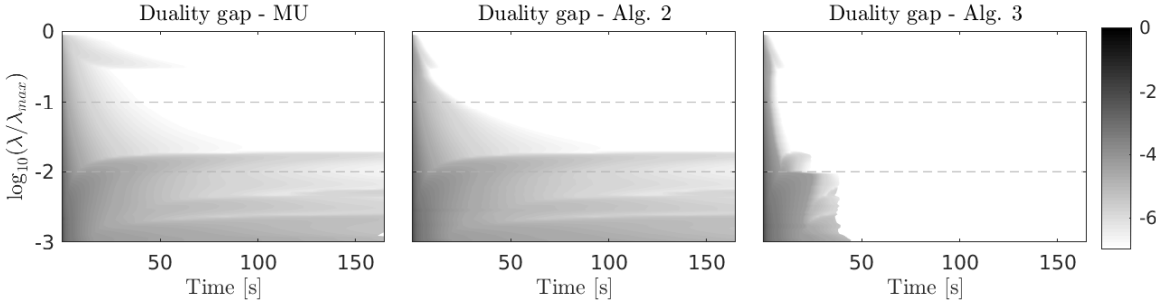
We here consider the NIPS papers word counts data set (Globerson et al., 2007). The input vector $\mathbf{y}$ is a randomly selected column of the data matrix. The remaining data forms matrix $\mathbf{A}$ (of size 2483×14035), after removing any all-zero rows and renormalizing all columns to unit-norm. Three standard optimization algorithms of distinct types have been used to solve



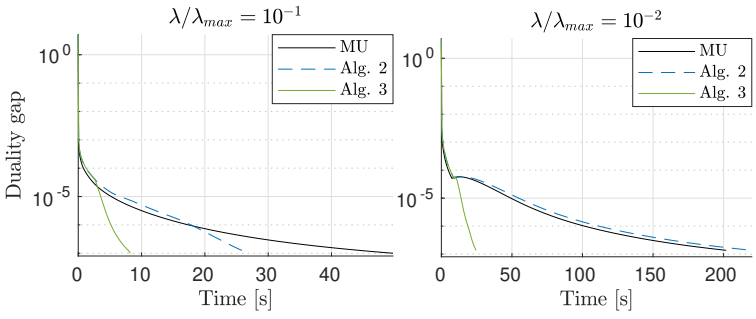
(a) Screening ratio against iterations for $\lambda/\lambda_{\max} \in [10^{-1}, 10^{-3}]$ (the lighter, the more screened coordinates) for Algorithms 2 and 3.



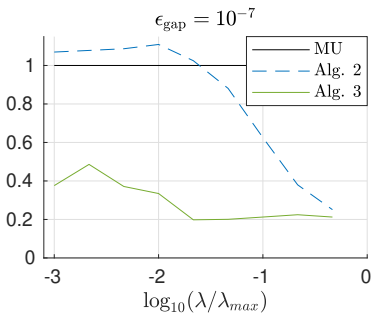
(b) Screening ratio against iterations for fixed regularization $\lambda/\lambda_{\max} = 10^{-1}$ (left) and $\lambda/\lambda_{\max} = 10^{-2}$ (right), corresponding to the dotted slices in Figure 4a.



(c) Convergence rate (duality gap) against execution time for $\lambda/\lambda_{\max} \in [10^{-1}, 10^{-3}]$ (the lighter, the closer to convergence). From left to right: MU solver alone and Algorithms 2 and 3.



(d) Duality gap against time for regularization $\lambda/\lambda_{\max} = 10^{-1}$ (left) and $\lambda/\lambda_{\max} = 10^{-2}$ (right), corresponding to the dotted slices in Figure 4c.



(e) Relative execution times for $\lambda/\lambda_{\max} \in [10^{-3}, 1]$.

Figure 4: Sparse non-negative hyperspectral decomposition using the Urban data set with $\beta = 1.5$ and MU.

	$\lambda/\lambda_{\max}$	10^{-1}		10^{-2}		10^{-3}	
	ε_{gap}	10^{-5}	10^{-7}	10^{-5}	10^{-7}	10^{-5}	10^{-7}
SPIRAL	(Dantas et al., 2021)	2.77	3.21	2.50	2.83	2.26	2.53
	Algorithm 2	8.81	9.61	8.68	9.69	8.54	9.36
	Algorithm 3	8.75	9.55	8.61	9.61	8.44	9.24
CoD	(Dantas et al., 2021)	4.19	5.35	4.06	5.13	4.12	5.06
	Algorithm 2	17.03	19.70	16.45	18.73	15.95	18.26
	Algorithm 3	17.68	20.57	16.38	18.66	16.18	18.51
MU	(Dantas et al., 2021)	6.71	8.88	6.67	9.52	5.74	7.31
	Algorithm 2	17.24	23.56	20.46	26.58	18.28	23.73
	Algorithm 2	16.56	24.76	19.17	24.89	17.40	22.42

Table 4: Sparse KL regression: Average speedups (ratio of execution times without and with screening) using the NIPS papers data set. Dantas et al. (2021) is equivalent to Alg. 2 but with a worse $\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$ constant.

problem (17): the multiplicative MM algorithm of (Lee and Seung, 2001; Févotte and Idier, 2011), coodinate descent (Hsieh and Dhillon, 2011) and proximal gradient descent (SPIRAL, Harmany et al., 2012). As the different solvers lead to qualitatively similar results, we have chosen to only report the results for the SPIRAL method in order to avoid redundancy. Yet, for completeness, speedup results for all solvers are summarized in Table 4.

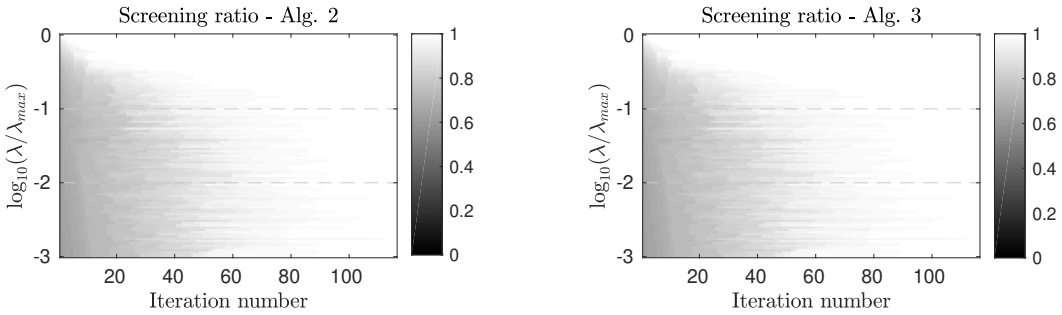
Differently from the previous cases, there is virtually no difference between Algorithms 2 and 3 here. This indicates that: 1) the strong concavity constant does not vary significantly within the dual feasible set while approaching the dual solution and 2) the initial bound for $\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$ given in Proposition 34 is nearly tight. This fact is verified both in terms of screening performance in Figures 5a and 5b and convergence time in Figures 5c and 5d.

Nonetheless, we still observe significant speedups w.r.t. the basic solver, which proves the interest of the proposed techniques (see Figures 5c and 5d and Table 4). Acceleration by a factor of 20 are reported in Table 4 with remarkably stable results across the different regularization regimes. The obtained results are also about 3 times better than those reported in our previous work (Dantas et al., 2021), which corresponds to Algorithm 2 with a different bound for $\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$. This indicates that the new strong concavity bound derived in the present paper is significantly tighter than the one in (Dantas et al., 2021). Indeed, a difference of around two orders of magnitude was observed experimentally between the two bounds.

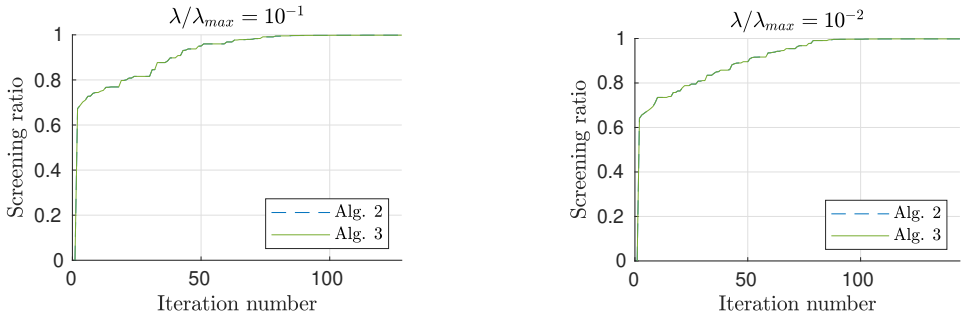
5.4 A Deeper Look

5.4.1 Strong concavity bound evolution and robustness to initialization

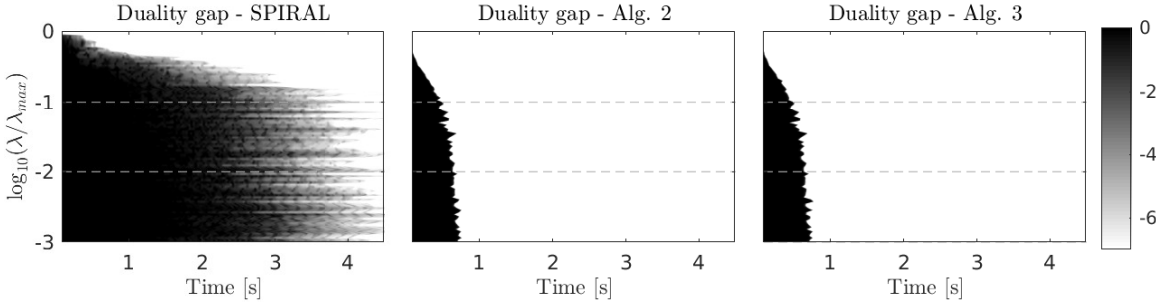
Figure 6 shows the value of the strong concavity bound α over the iterations in the logistic regression scenario. While it is kept constant in Algorithms 1 and 2, it is progressively refined in Algorithm 3. To evaluate the robustness of the proposed refinement approach, we run Algorithm 3 with two different initializations: the global constant $\alpha_{\mathbb{R}^m}$ and the local constant $\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$ (used respectively in Algorithms 1 and 2).



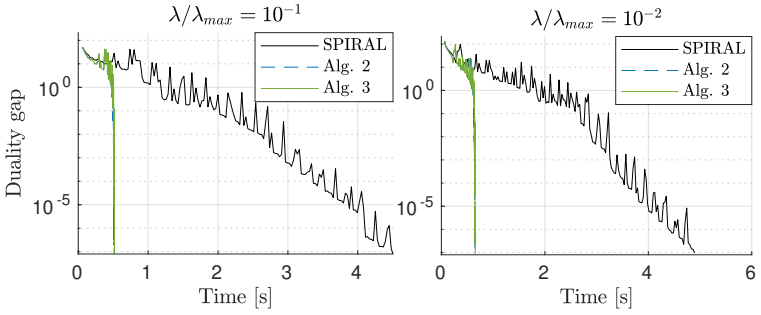
(a) Screening ratio against iterations for $\lambda/\lambda_{\max} \in [10^{-1}, 10^{-3}]$ (the lighter, the more screened coordinates), for Algorithms 2 and 3.



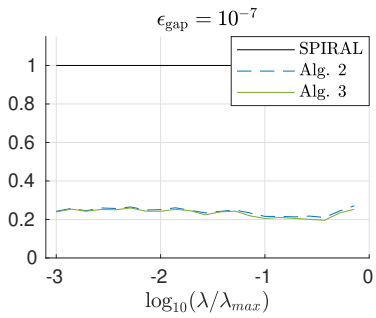
(b) Screening ratio against iterations for fixed regularization $\lambda/\lambda_{\max} = 10^{-1}$ (left) and $\lambda/\lambda_{\max} = 10^{-2}$ (right), corresponding to the dotted slices in Figure 5a.



(c) Convergence rate (duality gap) against execution time for $\lambda/\lambda_{\max} \in [10^{-1}, 10^{-3}]$ (the lighter, the closer to convergence). From left to right: SPIRAL solver alone and in Algorithms 2 and 3.



(d) Duality gap against time for regularization $\lambda/\lambda_{\max} = 10^{-1}$ (left) and $\lambda/\lambda_{\max} = 10^{-2}$ (right) corresponding to the dotted slices in Figure 5c.



(e) Relative execution times for $\lambda/\lambda_{\max} \in [10^{-3}, 1]$.

Figure 5: Sparse KL regression using NIPS papers word count data set and the SPIRAL solver.

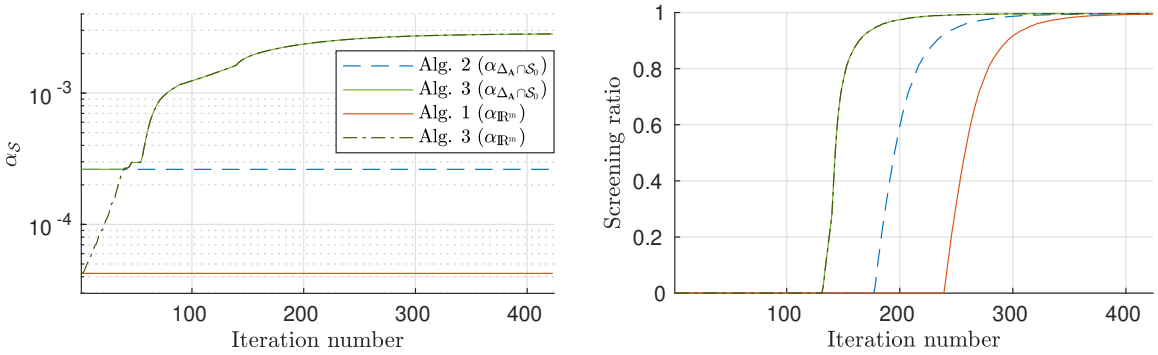


Figure 6: Left: evolution of the strong concavity bound α against iterations for the logistic regression problem, $\lambda/\lambda_{\max} = 10^{-3}$. Algorithm 3 is initialized with local $\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$ (solid light green line) and global $\alpha_{\mathbb{R}^m}$ (dash-dotted dark green line) bounds. Right: impact of the initialization over the screening performance.

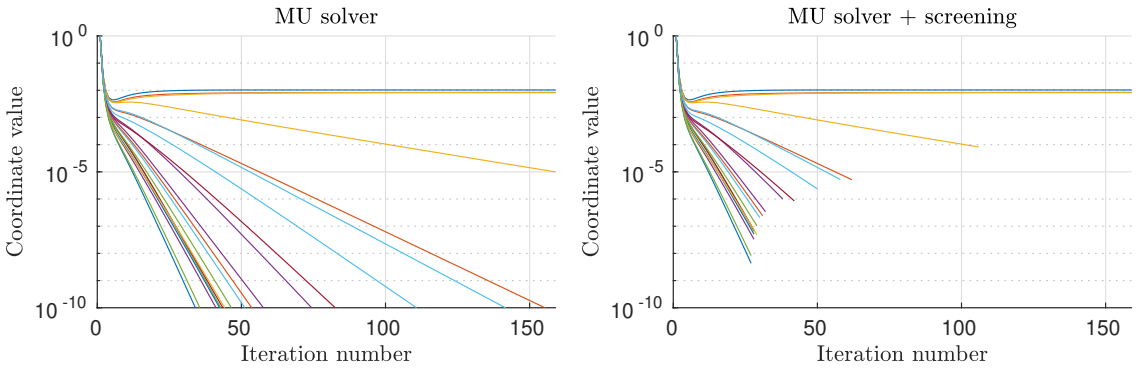
The left plot in Figure 6 shows that even with a poor initialization of α, the proposed refinement approach will quickly improve the provided bound to match the better initialization, even though there is about one order of magnitude difference between both initializations. This indicates that the proposed refinement approach makes Algorithm 3 quite robust to the initialization of α and that the global $\alpha_{\mathbb{R}^m}$ constant can be used as an initialization of Algorithm 3 instead of the proposed $\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$. This can be useful, for instance, when the additional full-rank assumption required by the latter bound is not met. More broadly, it suggests that Algorithm 3 can be applied to a new problem at hand without requiring an accurate initial estimation of the problem's strong-concavity constant (provided, obviously, that one is capable of performing the refinement step, i.e., computing the bounds over $\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0$, which tends to be easier).

In the example depicted in Figure 6 the poor initialization of α is compensated dozens of iterations before screening even starts to take place and, as a consequence, no performance loss is inflicted by the poor initialization. Obviously, in other cases, this compensation might not be as quick, causing some harm to the final execution time. Yet, in any case, the proposed refinement approach significantly mitigates the impact of a poor initialization on the overall screening performance (and execution time).

5.4.2 Support identification for MU solver

Finally, we discuss and illustrate the power of screening for MU solvers using a small-size synthetic experiment (for readability). Multiplicative updates are very standard in (sparse) NMF, in particular with the general β-divergence (Févotte and Idier, 2011)[8] but suffer from a well-known limitation. Because each coordinate (either of the dictionary or activation matrix) is multiplied by a strictly positive factor, convergence to zero values can only be asymptotical. This is shown in the left plot of Figure 7 which shows the value of each coordinate x_j of the primal estimate $\mathbf{x}$ over the iterations for the MU solver on the $\beta_{1.5}$-divergence case with

8. More efficient, e.g., proximal-based, coordinate-descent or active set methods exist for the quadratic (Friedman et al., 2010; Beck and Teboulle, 2009; Johnson and Guestrin, 2015) or KL particular cases (Harmany et al., 2012; Hsieh and Dhillon, 2011; Virtanen et al., 2013).



(a) (b)

Figure 7: Coordinate values over the iterations of an MU solver (a) without screening and (b) with screening. Each colored line corresponds to one coordinate, the line stops when the coordinate is screened.

dimensions $m = 10$ and $n = 20$. In practice, some arbitrary thresholding may be performed to force some entries to zero. However, such an ad-hoc operation may mistakenly cancel coordinates that belong to the support but happen to have small values. Conversely, the value of some coordinates might decrease very slowly with the iterations (see for instance the yellow line in Figure 7a) and may be mistakenly kept by such an arbitrary thresholding procedure. The proposed screening approach (Figure 7b) tackles this issue by introducing *actual* zeros to the solution with theoretical guarantees. Application of such strategies in NMF settings is an exciting and potentially fruitful perspective of this work.

6. Conclusion

In this paper, we proposed a safe screening framework that improves upon the existing Gap Safe screening approach, while extending its application to a wider range of problems—in particular, problems whose associated dual problem is not globally strongly concave.

Two screening algorithms have been proposed in this new framework, exploiting local properties of the involved functions. First, we define a direct extension of the conventional dynamic screening approach, by replacing the global strong concavity bound $\alpha_{\mathbb{R}^m}$ by a local one $\alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$ in Algorithm 2. Noting that a reinforcement loop arises between the current safe sphere and the strong concavity bound within the sphere itself, we propose the iterative refinement approach in Algorithm 3. By construction, the latter approach can only improve the initial strong-concavity bound, which makes it strictly superior to the former (the computational overhead due to the extra refinement loop was empirically observed to be negligible). Both proposed approaches lead to considerable speedups on several existing solvers and for various simulation scenarios.

Algorithm 3 can be superior to Algorithm 2 for two main reasons: 1) The strong concavity constant varies significantly within the dual feasible set. Therefore, refining it on the Gap Safe sphere (which shrinks over the iterations) may lead to significant improvements. 2) The initial strong concavity bound is loose (for instance, because it cannot be computed exactly

and no tighter estimation is available). This initial handicap will then be progressively compensated by the refinement procedure in Algorithm 3.

The proposed framework is quite generic and not restricted to the four treated cases. Its application to other problems demands the completion of the following few steps: 1) the possible definition of a subset $\mathcal{S}_0$ (only for most challenging objective functions whose dual is not strongly-concave on the entire dual feasible set); 2) the computation of an initial strong concavity bound α such that $\alpha \leq \alpha_{\Delta_{\mathbf{A}} \cap \mathcal{S}_0}$ (as observed experimentally, this bound does not need to be very precise); 3) the ability to compute the strong concavity bound on a ℓ_2-ball (which possibly intersects with $\mathcal{S}_0$) for the refinement step.

Although we restricted ourselves the dynamic screening setting, nothing prevents this idea to be applied to more advanced Gap Safe screening configurations: combined with sequential screening (Ndiaye et al., 2017), working sets (Massias et al., 2017), dual extrapolation techniques (Massias et al., 2020), or even the stable safe screening framework (Dantas and Gribonval, 2019) in which approximation errors are tolerated in the data matrix.

Acknowledgments

This work was supported by the European Research Council (ERC FACTORY-CoG-6681839).

Appendices

Table of Contents

Appendix A. Dual Problem and Optimality Conditions

A.1 Preliminaries

In order to prove Theorem 1 and its ensuing particular cases, we will a the classical result in (Borwein and Lewis, 2000, Theorem 3.3.5) (Rockafellar, 1970, Theorem 31.3) which relates generic primal and dual problems with the Fenchel conjugates of the composing functions.

Generic Primal. Consider a generic primal problem of the form with F (resp G) a closed proper and convex function on $\mathbb{R}^m$ (resp. $\mathbb{R}^n$):

$$\mathbf{x}^\star = \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}} \underbrace{F(\mathbf{A}\mathbf{x}) + \lambda G(\mathbf{x})}_{:=P_\lambda(\mathbf{x})} \tag{35}$$

where we denote $P_\lambda(\mathbf{x})$ the primal cost function.

Generic Dual. The associated Lagrangian dual problem can be shown to be given by Borwein and Lewis (2000, Theorem 3.3.5)

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}\in\mathbb{R}^m}{\operatorname{argmax}} \underbrace{-F^*(-\lambda\boldsymbol{\theta}) - \lambda G^*(A^\mathsf{T}\boldsymbol{\theta})}_{=:D_\lambda(\boldsymbol{\theta})} \tag{36}$$

where $D_\lambda(\boldsymbol{\theta})$ denotes the dual (cost) function.

The optimality conditions are given by the following result (Bauschke and Combettes, 2011, Theorem 19.1).

Theorem 12. *Let $F : \mathbb{R}^m \to (-\infty, +\infty]$ and $G : \mathbb{R}^n \to (-\infty, +\infty]$ be convex lower semi-continuous, and let the linear operator $\mathbf{A} \in \mathbb{R}^{m\times n}$ be such that $\operatorname{dom}(G) \cap \mathbf{A}\operatorname{dom}(F) \neq \emptyset$. Then the following are equivalent:*

(i) $\mathbf{x}^\star$ is a primal solution, $\boldsymbol{\theta}^\star$ is a dual solution and $P_\lambda(\mathbf{x}^\star) = D_\lambda(\boldsymbol{\theta}^\star)$ (i.e. strong duality holds).

(ii) $\mathbf{A}^\mathsf{T}\boldsymbol{\theta}^\star \in \partial G(\mathbf{x}^\star)$ and $-\lambda\boldsymbol{\theta}^\star \in \partial F(\mathbf{A}\mathbf{x}^\star)$.

Other useful results are given below:

Proposition 13 (Separable sum property of the Fenchel conjugate)**.** *(Hiriart-Urruty and Lemaréchal, 1993b, Ch. X, Proposition 1.3.1 (ix)) Let $F = \sum_{i=1}^m f_i$ be coordinate-wise separable, then*

$$F^*(\mathbf{u}) = \sum_{i=1}^m f_i^*(u_i)$$

Proposition 14. *Let $F(\mathbf{z}) = \sum_{i=1}^m f_i(z_i)$ be differentiable, then its gradient $\nabla F : \mathbb{R}^m \to \mathbb{R}^m$ is given by the (scalar) derivatives of f_i as follows*

$$\nabla F(\mathbf{z}) = [f_1'(z_1), \ldots, f_m'(z_m)]^\mathsf{T} \in \mathbb{R}^m.$$

Proposition 15 (Dual norm of a group-decomposable norm)**.** *(Ndiaye, 2018, Proposition 21) Let $\mathcal{G}$ be a partition of $[n]$ and $\Omega(\mathbf{u}) = \sum_{g\in\mathcal{G}} \Omega_g(\mathbf{u}_g)$ be a group-decomposable norm. Its dual norm is given by $\overline{\Omega}(\mathbf{u}) = \max_{g\in\mathcal{G}} \overline{\Omega}_g(\mathbf{u}_g)$ where $\overline{\Omega}_g$ is the dual norm of Ω_g*

Proof The result follows from Proposition 13: $\Omega^*(\mathbf{u}) = \sum_{g\in\mathcal{G}} \Omega_g^*(\mathbf{u}_g) = \sum_{g\in\mathcal{G}} \mathbb{1}_{\overline{\Omega}_g(\mathbf{u}_g)\leq 1} = \mathbb{1}_{\{\mathbf{u} \mid \forall g\in\mathcal{G},\, \overline{\Omega}_g(\mathbf{u}_g)\leq 1\}} = \mathbb{1}_{\max_{g\in\mathcal{G}}(\overline{\Omega}_g(\mathbf{u}_g))\leq 1}$. Since $\Omega^*(\cdot) = \mathbb{1}_{\overline{\Omega}(\cdot)}$ we conclude that $\overline{\Omega}(\mathbf{u}) = \max_{g\in\mathcal{G}} \big(\overline{\Omega}_g(\mathbf{u}_g)\big)$, which finishes the proof. ■

A.2 Proof of Theorem 1

Note that in formulation (35), there are no constraints on the primal variable $\mathbf{x}$. To make Problem (1) fit to this setting, we reformulate it as an unconstrained problem by using the indicator function $\mathbb{1}_\mathcal{C}(\mathbf{x})$. This strategy was used in Wang et al. (2019) for the non-negative Lasso problem and is extended here to a boarder class of problems.

$$\mathbf{x}^\star \in \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^m f_i([\mathbf{A}\mathbf{x}]_i) + \lambda\left(\Omega(\mathbf{x}) + \mathbb{1}_\mathcal{C}(\mathbf{x})\right) \tag{37}$$

Comparing (37) to (35), we have the following direct correspondences:

- $F(\mathbf{Ax}) = \sum_{i=1}^{m} f_i([\mathbf{Ax}]_i)$
- $G(\mathbf{x}) = G_1(\mathbf{x}) + G_2(\mathbf{x}) = \Omega(\mathbf{x}) + \mathbb{1}_{\mathcal{C}}(\mathbf{x})$

A.2.1 Dual problem

Then, the dual problem in (4) is a direct application of the generic result in (36). We just need to evaluate the Fenchel conjugates F^* and G^* in our particular case described above.

$F = \sum_{i=1}^{m} f_i$ being coordinate-wise separable, we can apply Proposition 13:

$$F(\mathbf{z}) = \sum_{i=1}^{m} f_i(z_i) \;\Rightarrow\; F^*(\mathbf{u}) = \sum_{i=1}^{m} f_i^*(u_i) \tag{38}$$

Furthermore, we know that the Fenchel conjugate of a norm Ω is the indicator on the unit-ball of its corresponding dual norm $\overline{\Omega}$ (see for instance Bauschke and Combettes, 2011, Example 13.32):

$$G_1(\mathbf{x}) = \Omega(\mathbf{x}) \quad \Rightarrow \quad G_1^*(\mathbf{u}) = \mathbb{1}_{\overline{\Omega}(\mathbf{u})\leq 1}$$

To conclude the proof, we need to consider the indicator function corresponding to constraint set $\mathcal{C}$. The case $\mathcal{C} = \mathbb{R}^n$ is trivial since $\mathbb{1}_{\mathbb{R}^n}(\mathbf{x}) \equiv 0$ and the Fenchel conjugate $G^* = G_1^*$ is given above. For the non-negativity contraint, $\mathcal{C} = \mathbb{R}^n_+$, we have (Wang et al., 2019, Lemma 18 (i)):

$$G_2(\mathbf{x}) = \mathbb{1}_{\mathbb{R}^n_+}(\mathbf{x}) \quad \Rightarrow \quad G_2^*(\mathbf{u}) = \mathbb{1}_{\mathbb{R}^n_-}(\mathbf{u}). \tag{39}$$

Now, to calculate the conjugate of the sum $\Omega(\mathbf{x}) + \mathbb{1}_{\mathbb{R}^n_+}(\mathbf{x})$, we use a property that relates the conjugate of a sum to the so-called *infimal convolution*, denoted $\boxdot$, of the individual conjugates (Bauschke and Combettes, 2011, Proposition 15.2).

$$G^*(\mathbf{u}) = (G_1 + G_2)^*(\mathbf{u}) = (G_1^* \boxdot G_2^*)(\mathbf{u}) = \mathbb{1}_{\overline{\Omega}(\mathbf{u}\leq 1)} \boxdot \mathbb{1}_{\mathbb{R}^n_-}(\mathbf{u}) \tag{40}$$

Then, we use the fact that the infimal convolution of two indicator functions is the indicator of Minkowski sum of both sets (Bauschke and Combettes, 2011, Example 12.3):

$$G^*(\mathbf{u}) = \mathbb{1}_{\overline{\Omega}(\mathbf{u})\leq 1} \boxdot \mathbb{1}_{\mathbb{R}^n_-}(\mathbf{u}) = \mathbb{1}_{\{\mathbf{u}=\mathbf{a}+\mathbf{b} \;|\; \overline{\Omega}(\mathbf{a})\leq 1,\; \mathbf{b}\leq 0\}} = \mathbb{1}_{\overline{\Omega}([\mathbf{u}]^+)\leq 1} \tag{41}$$

Remark 16. *The last step applies only because both conjugates G_1^* and G_2^* are indicator functions. This is no longer the case when we assume, for instance, a bounded-variable constraint of the form $x_j \in [-a_j, b_j],\ j \in [n]$.*

This leads to the following results:

- For $\mathcal{C} = \mathbb{R}^n$, we have $G^*(\mathbf{u}) = G_1^*(\mathbf{u}) = \mathbb{1}_{\overline{\Omega}(\mathbf{u})\leq 1}$.
- For $\mathcal{C} = \mathbb{R}^n_+$, we have $G^*(\mathbf{u}) = (G_1 + G_2)^*(\mathbf{u}) = \mathbb{1}_{\overline{\Omega}([\mathbf{u}]^+)\leq 1}$.

which can be written in a compact form using the non-linearity function ϕ as defined in (6):

$$G^*(\mathbf{u}) = \mathbb{1}_{\overline{\Omega}(\phi(\mathbf{u}))\leq 1} \tag{42}$$

Applying Proposition 15 for the dual of a group-separable norm $\overline{\Omega}(\phi(\mathbf{u})) = \max_{g\in\mathcal{G}}\big(\overline{\Omega}_g(\phi(\mathbf{u}_g))\big)$:

$$G^*(\mathbf{u}) = \mathbb{1}_{\max_{g\in\mathcal{G}}\left(\overline{\Omega}_g(\phi(\mathbf{u}_g))\right)\leq 1}. \tag{43}$$

The resulting dual problem is thus obtained by replacing F^* (38) and G^* (43) in the generic dual problem (36), that is

$$D_\lambda(\boldsymbol{\theta}) = -F^*(-\lambda\boldsymbol{\theta}) - \mathbb{1}_{\overline{\Omega}(\phi(\mathbf{A}^\top\boldsymbol{\theta}))\leq 1}. \tag{44}$$

Remark 17. *The constraint set being coordinate separable, we can always treat each group separately by restricting ourselves to the set of n_g coordinates on group g. For simplicity, we derive the remaining our results for Ω on $\mathbb{R}^n$ (the extension for groups with Ω_g being straightforward).*

A.2.2 Optimality conditions

Optimality conditions (7) and (8) are a direct application of Theorem 12. Applying our definitions of F and G, and knowing that $\partial F(\mathbf{z}) = \{\nabla F(\mathbf{z})\}$ (a singleton as, by assumption, F is differentiable) we have:

$$-\lambda\boldsymbol{\theta}^\star = \nabla F(\mathbf{A}\mathbf{x}^\star) = \big[f_1'([\mathbf{A}\mathbf{x}^\star]_1), \ldots, f_m'([\mathbf{A}\mathbf{x}^\star]_m)\big]^\mathsf{T} \tag{45}$$

$$\mathbf{A}^\mathsf{T}\boldsymbol{\theta}^\star \in \partial G(\mathbf{x}^\star) = \partial\Omega(\mathbf{x}^\star) + \partial\mathbb{1}_{\mathcal{C}}(\mathbf{x}^\star) \tag{46}$$

To obtain (8) explicitly, the sum of subdifferentials in equation (46) is further developed below.

A.2.3 Sum of subdifferentials: $\partial\Omega(\mathbf{x}) + \partial\mathbb{1}_{\mathcal{C}}(\mathbf{x})$

Let us further develop the second optimality condition (46). It requires computing the subdifferentials of both Ω and $\mathbb{1}_{\mathcal{C}}$, which are discussed below:

Proposition 18 (Subdifferential of a Norm)**.** *(Bach et al., 2012, Proposition 1.2) The subdifferential of a norm $\Omega : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{x}$ is given by*

$$\partial\Omega(\mathbf{x}) = \begin{cases} \{\mathbf{z}\in\mathbb{R}^n \mid \overline{\Omega}(\mathbf{z}) \leq 1\}, & \text{if } \mathbf{x} = \mathbf{0} \\ \{\mathbf{z}\in\mathbb{R}^n \mid \overline{\Omega}(\mathbf{z}) = 1 \text{ and } \mathbf{z}^\mathsf{T}\mathbf{x} = \Omega(\mathbf{x})\}, & \text{otherwise.} \end{cases} \tag{47}$$

Proposition 19 (Subdifferential of an indicator function)**.** *(Bauschke and Combettes, 2011, Example 16.13) The subdifferential at point $\mathbf{x}$ of the indicator of a set $\mathcal{C}$ is given by the normal cone of $\mathcal{C}$ at $\mathbf{x}$, denoted $N_{\mathcal{C}}(\mathbf{x})$:*

$$\partial\mathbb{1}_{\mathcal{C}}(\mathbf{x}) = N_{\mathcal{C}}(\mathbf{x}) = \begin{cases} \{\mathbf{z}\in\mathbb{R}^n \mid \mathbf{z}^\mathsf{T}\mathbf{x} \geq \mathbf{z}^\mathsf{T}\mathbf{w},\ \forall\mathbf{w}\in\mathcal{C}\} & \text{if } \mathbf{x}\in\mathcal{C} \\ \emptyset & \text{otherwise.} \end{cases} \tag{48}$$

In particular, $N_{\mathcal{C}}(\mathbf{x}) = \{\mathbf{0}\}, \forall\mathbf{x}\in\operatorname{int}(\mathcal{C})$.

When it comes to the constraint sets considered in this paper, we have:

$$\partial \mathbb{1}_{\mathbb{R}^n} \equiv \{0\}. \tag{49}$$

$$\partial \mathbb{1}_{R_+^n}(\mathbf{x}) = J_1 \times \cdots \times J_n, \qquad \text{with } J_j = \begin{cases} \emptyset & \text{if } x_j < 0 \\ (-\infty, 0] & \text{if } x_j = 0 \\ \{0\} & \text{if } x_j > 0. \end{cases} \tag{50}$$

The case $\mathcal{C} = \mathbb{R}^n$ is trivial, since $\partial \mathbb{1}_{\mathbb{R}^n}(\mathbf{x}) \equiv \{\mathbf{0}\}$ and therefore $\partial\Omega(\mathbf{x}) + \partial \mathbb{1}_{\mathbb{R}^n}(\mathbf{x}) = \partial\Omega(\mathbf{x})$, which is given in equation (47). Let us now analyze the case $\mathcal{C} = \mathbb{R}^n_+$. First of all, note that the subdifferential $\partial \mathbb{1}_{\mathbb{R}^n_+}(\mathbf{x})$ in equation (50) can be rewritten in the following form:

$$\partial \mathbb{1}_{R_+^n}(\mathbf{x}) = \begin{cases} \emptyset & \text{if } \mathbf{x} \notin \mathbb{R}^n_+ \\ \{\mathbf{z} \in \mathbb{R}^n \mid \mathbf{z} \leq \mathbf{0}\} = \mathbb{R}^n_- & \text{if } \mathbf{x} = \mathbf{0} \\ \{\mathbf{z} \in \mathbb{R}^n \mid \mathbf{z} \leq \mathbf{0} \text{ and } \mathbf{z}^\mathsf{T}\mathbf{x} = 0\} & \text{otherwise} \quad (\text{i.e. } \mathbf{x} \geq \mathbf{0} \text{ and } \mathbf{x} \neq \mathbf{0}). \end{cases} \tag{51}$$

Therefore, the (Minkowski) sum of the subdifferential sets in equations (47) and (51) gives:

$$\begin{aligned} \partial\Omega(\mathbf{x}) + \partial \mathbb{1}_{\mathbb{R}^n_+}(\mathbf{x}) &= \begin{cases} \emptyset & \text{if } \mathbf{x} \notin \mathbb{R}^n_+ \\ \{\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_2 \mid \overline{\Omega}(\mathbf{z}_1) \leq 1,\ \mathbf{z}_2 \leq 0\}, & \text{if } \mathbf{x} = \mathbf{0} \\ \{\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_2 \mid \overline{\Omega}(\mathbf{z}_1) = 1,\ \mathbf{z}_1^\mathsf{T}\mathbf{x} = \Omega(\mathbf{x}), \mathbf{z}_2 \leq 0,\ \mathbf{z}_2^\mathsf{T}\mathbf{x} = 0\}, & \text{otherwise} \end{cases} \\ &= \begin{cases} \emptyset & \text{if } \mathbf{x} \notin \mathbb{R}^n_+ \\ \{\mathbf{z} \in \mathbb{R}^n \mid \overline{\Omega}([\mathbf{z}]^+) \leq 1\}, & \text{if } \mathbf{x} = \mathbf{0} \\ \{\mathbf{z} \in \mathbb{R}^n \mid \overline{\Omega}([\mathbf{z}]^+) = 1,\ \mathbf{z}^\mathsf{T}\mathbf{x} = \Omega(\mathbf{x})\}, & \text{otherwise} \end{cases} \end{aligned} \tag{52}$$

where, in the last equality, we have used the fact that $\mathbf{z}_1^\mathsf{T}\mathbf{x} = (\mathbf{z} - \mathbf{z}_2)^\mathsf{T}\mathbf{x} = \mathbf{z}^\mathsf{T}\mathbf{x}$, since $\mathbf{z}_2^\mathsf{T}\mathbf{x} = 0$.

Comparing equations (47) and (52) respectively for $\mathcal{C} = \mathbb{R}^n$ and $\mathcal{C} = \mathbb{R}^n_+$, we can see that the sum $\partial\Omega + \partial \mathbb{1}_{\mathcal{C}}$ for the considered constraint sets writes compactly as:

$$\partial\Omega(\mathbf{x}) + \partial \mathbb{1}_{\mathcal{C}}(\mathbf{x}) = \begin{cases} \emptyset & \text{if } \mathbf{x} \notin \mathcal{C} \\ \{\mathbf{z} \in \mathbb{R}^n \mid \overline{\Omega}(\phi(\mathbf{z})) \leq 1\}, & \text{if } \mathbf{x} = \mathbf{0} \\ \{\mathbf{z} \in \mathbb{R}^n \mid \overline{\Omega}(\phi(\mathbf{z})) = 1,\ \mathbf{z}^\mathsf{T}\mathbf{x} = \Omega(\mathbf{x})\}, & \text{otherwise} \end{cases} \tag{53}$$

where ϕ is defined as in (6).

Applying the above results to equation (46), it takes the following form which corresponds to (8)

$$\begin{cases} \overline{\Omega}(\phi(\mathbf{A}^\mathsf{T}\boldsymbol{\theta}^\star)) \leq 1, & \text{if } \mathbf{x}^\star = \mathbf{0} \\ \overline{\Omega}(\phi(\mathbf{A}^\mathsf{T}\boldsymbol{\theta}^\star)) = 1, \quad (\mathbf{A}^\mathsf{T}\boldsymbol{\theta}^\star)^\mathsf{T}\mathbf{x}^\star = \Omega(\mathbf{x}^\star), & \text{otherwise} \end{cases} \tag{54}$$

Appendix B. Proof of Theorem 5

In this section we prove Theorem 5, i.e. given any feasible primal-dual pair $(\mathbf{x}, \boldsymbol{\theta}) \in (\mathrm{dom}(P_\lambda) \cap \mathcal{C}) \times (\Delta_A \cap \mathcal{S})$ and supposing D_λ to be α_S-strongly concave on $\mathcal{S}$, we prove that a region

$$\mathcal{B}(\boldsymbol{\theta}, r), \quad \text{with } r = \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha_\mathcal{S}}} \tag{55}$$

is safe, i.e. $\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r)$. Equivalently, we prove that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2 \leq r$.

The proof follows quite similarly to Ndiaye et al. (2017, Theorem 6). (except that we assume strong concavity of D_λ exclusively on $\mathcal{S}$ instead of $\mathbb{R}^m$). From the $\alpha_\mathcal{S}$-strong concavity of D_λ on $\mathcal{S}$ we have:

$$\forall(\boldsymbol{\theta}', \boldsymbol{\theta}'') \in S \times S, \quad D_\lambda(\boldsymbol{\theta}'') \leq D_\lambda(\boldsymbol{\theta}') + \langle \nabla D_\lambda(\boldsymbol{\theta}'), \boldsymbol{\theta}'' - \boldsymbol{\theta}' \rangle - \frac{\alpha_S}{2}\|\boldsymbol{\theta}'' - \boldsymbol{\theta}'\|_2^2.$$

In particular, we can take $\boldsymbol{\theta}' = \boldsymbol{\theta}^\star$, $\boldsymbol{\theta}'' = \boldsymbol{\theta}$ (since we suppose $\boldsymbol{\theta} \in \mathcal{S}$ and $\boldsymbol{\theta}^\star \in \mathcal{S}$):

$$D_\lambda(\boldsymbol{\theta}) \leq D_\lambda(\boldsymbol{\theta}^\star) + \langle \nabla D_\lambda(\boldsymbol{\theta}^\star), \boldsymbol{\theta} - \boldsymbol{\theta}^\star \rangle - \frac{\alpha_S}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2.$$

By definition $\boldsymbol{\theta}^\star$ maximizes D_λ on $\Delta_\mathbf{A}$ and hence $\langle \nabla D_\lambda(\boldsymbol{\theta}^\star), \boldsymbol{\theta} - \boldsymbol{\theta}^\star \rangle \leq 0$ (otherwise, $\boldsymbol{\theta} - \boldsymbol{\theta}^\star$ would be a feasible direction of improvement), which implies:

$$D_\lambda(\boldsymbol{\theta}) \leq D_\lambda(\boldsymbol{\theta}^\star) - \frac{\alpha_S}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2.$$

By weak duality $D_\lambda(\boldsymbol{\theta}^\star) \leq P_\lambda(\mathbf{x}')$, $\forall \mathbf{x}' \in \mathrm{dom}(P_\lambda) \cap \mathcal{C}$, and in particular for $\mathbf{x}' = \mathbf{x}$, giving:

$$\begin{aligned} D_\lambda(\boldsymbol{\theta}) &\leq P_\lambda(\mathbf{x}) - \frac{\alpha_S}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2 \\ \underbrace{D_\lambda(\boldsymbol{\theta}) - P_\lambda(\mathbf{x})}_{-G_\lambda(\mathbf{x}, \boldsymbol{\theta})} &\leq -\frac{\alpha_S}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2 \\ \frac{2}{\alpha_S} G_\lambda(\mathbf{x}, \boldsymbol{\theta}) &\geq \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2 \end{aligned}$$

which holds for all $\mathbf{x} \in \mathrm{dom}(P_\lambda) \cap \mathcal{C}, \boldsymbol{\theta} \in \Delta_\mathbf{A} \cap \mathcal{S}$, concluding the proof.

Appendix C. Proofs of Section 4.1

C.1 Proof of Proposition 8

Statement.

$$\mathbf{0} \in \operatorname*{argmin}_{\mathbf{x} \in \mathcal{C}} P_\lambda(\mathbf{x}) \iff \lambda \geq \lambda_{\max} := \overline{\Omega}\left(\phi\left(-\mathbf{A}^\mathsf{T} \nabla F(\mathbf{0})\right)\right) \tag{56}$$

Proof Writing our primal problem in an equivalent unconstrained form, we have

$$\operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} F(\mathbf{A}\mathbf{x}) + \lambda\left(\Omega(\mathbf{x}) + \mathbb{1}_\mathcal{C}(\mathbf{x})\right) := P_\lambda^{\mathrm{unc}}(\mathbf{x})$$

denoting P_λ^{unc} the unconstrained objective function, with $\operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^n} P_\lambda^{\mathrm{unc}}(\mathbf{x}) = \operatorname{argmin}_{\mathbf{x} \in \mathcal{C}} P_\lambda(\mathbf{x})$. From Fermat's rule we have that

$$\begin{aligned} \mathbf{0} \in \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} P_\lambda^{\mathrm{unc}}(\mathbf{x}) &\iff \mathbf{0} \in \partial P_\lambda^{\mathrm{unc}}(\mathbf{0}) = \{\mathbf{A}^\mathsf{T} \nabla F(\mathbf{0})\} + \lambda\left(\partial\Omega(\mathbf{0}) + \partial \mathbb{1}_\mathcal{C}(\mathbf{0})\right) \\ &\iff -\mathbf{A}^\mathsf{T} \nabla F(\mathbf{0})/\lambda \in \partial\Omega(\mathbf{0}) + \partial \mathbb{1}_\mathcal{C}(\mathbf{0}). \end{aligned}$$

Using the expression of $\partial\Omega(\mathbf{x}) + \partial\mathbb{1}_\mathcal{C}(\mathbf{x})$ given in (53) we obtain

$$\mathbf{0} \in \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} P_\lambda^{\mathrm{unc}}(\mathbf{x}) \iff \overline{\Omega}\left(\phi(-\mathbf{A}^\mathsf{T} \nabla F(\mathbf{0}))\right) \leq \lambda,$$

which completes the proof. ■

C.2 Proof of Lemma 9

Statement. Assume that $\mathrm{dom}(D_\lambda)$ is stable by contraction and let $\boldsymbol{\Xi} : \mathbb{R}^m \to \mathbb{R}^m$ be the scaling operator defined by

$$\boldsymbol{\Xi}(\mathbf{z}) := \frac{\mathbf{z}}{\max\left(\overline{\Omega}(\phi(\mathbf{A}^\mathsf{T}\mathbf{z})), 1\right)}. \tag{57}$$

Then, for any point $\mathbf{z} \in \mathrm{dom}(D_\lambda)$, we have $\boldsymbol{\Xi}(\mathbf{z}) \in \Delta_\mathbf{A}$. Moreover, for any primal point $\mathbf{x} \in \mathrm{dom}(P_\lambda)$, we have that $\mathbf{z} = (-\nabla F(\mathbf{A}\mathbf{x})/\lambda) \in \mathrm{dom}(D_\lambda)$ and therefore

$$\boldsymbol{\Xi}(-\nabla F(\mathbf{A}\mathbf{x})/\lambda) \in \Delta_\mathbf{A}. \tag{58}$$

Finally, if $F \in C^1$, then $\boldsymbol{\Xi}(-\nabla F(\mathbf{A}\mathbf{x})/\lambda) \to \boldsymbol{\theta}^\star$ as $\mathbf{x} \to \mathbf{x}^\star$.

Proof By definition of $\boldsymbol{\Xi}$, one can see that $\boldsymbol{\Xi}(\mathbf{z}) \leq \mathbf{z}$ and, given the assumption that $\mathrm{dom}(D_\lambda)$ is stable by contraction, we get that $\forall \mathbf{z} \in \mathrm{dom}(D_\lambda)$, $\boldsymbol{\Xi}(\mathbf{z}) \in \mathrm{dom}(D_\lambda)$. Combining this with the fact that $\overline{\Omega}\left(\phi(\mathbf{A}^\mathsf{T}\boldsymbol{\Xi}(\mathbf{z}))\right) \leq 1$ (by definition of $\boldsymbol{\Xi}$), we obtain

$$\forall \mathbf{z} \in \mathrm{dom}(D_\lambda),\ \boldsymbol{\Xi}(\mathbf{z}) \in \Delta_\mathbf{A}, \tag{59}$$

where we recall that $\Delta_\mathbf{A} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \overline{\Omega}(\phi(\mathbf{A}^\mathsf{T}\boldsymbol{\theta})) \leq 1\} \cap \mathrm{dom}(D_\lambda)$.

To complete the proof, it remains to show that, for any primal point $\mathbf{x} \in \mathrm{dom}(P_\lambda)$, $(-\nabla F(\mathbf{A}\mathbf{x})/\lambda) \in \mathrm{dom}(D_\lambda)$. Because $D_\lambda = -F^*(-\lambda\cdot)$, we get that $\boldsymbol{\theta} \in \mathrm{dom}(D_\lambda) \iff -\lambda\boldsymbol{\theta} \in \mathrm{dom}(F^*)$. Now, let $\mathbf{x} \in \mathrm{dom}(P_\lambda)$, then

$$F^*(\nabla F(\mathbf{A}\mathbf{x})) = \sup_{\mathbf{z}\in\mathbb{R}^m} \langle \mathbf{z}, \nabla F(\mathbf{A}\mathbf{x})\rangle - F(\mathbf{z}) = \langle \mathbf{A}\mathbf{x}, \nabla F(\mathbf{A}\mathbf{x})\rangle - F(\mathbf{A}\mathbf{x}) < \infty. \tag{60}$$

(The sup is attained for vector(s) $\mathbf{z} \in \mathbb{R}^m$ such that $\nabla F(\mathbf{A}\mathbf{x}) - \nabla F(\mathbf{z}) = 0$, such as $\mathbf{z} = \mathbf{A}\mathbf{x}$.) This shows that $\nabla F(\mathbf{A}\mathbf{x}) \in \mathrm{dom}(F^*)$ and thus that $(-\nabla F(\mathbf{A}\mathbf{x})/\lambda) \in \mathrm{dom}(D_\lambda)$.

Lastly, assuming that $F \in C^1$, we get by continuity of $\nabla F(\mathbf{A}\mathbf{x})$ that $\nabla F(\mathbf{A}\mathbf{x}) \to \nabla F(\mathbf{A}\mathbf{x}^\star)$ as $\mathbf{x} \to \mathbf{x}^\star$ and, from the primal-dual link in optimality condition (7) we have that $-\nabla F(\mathbf{A}\mathbf{x}^\star) = \lambda\boldsymbol{\theta}^\star$. Then, because $\boldsymbol{\theta}^\star \in \Delta_\mathbf{A}$, the scaling factor in $\boldsymbol{\Xi}$ is equal to 1 (by definition) which leads to $\boldsymbol{\Xi}(-\nabla F(\mathbf{A}\mathbf{x}^\star)/\lambda) = -\nabla F(\mathbf{A}\mathbf{x}^\star)/\lambda = \boldsymbol{\theta}^\star$. Finally, by continuity of $\boldsymbol{\Xi}$ (which is a composition of continuous functions), we conclude that $\boldsymbol{\Xi}(-\nabla F(\mathbf{A}\mathbf{x})/\lambda) \to \boldsymbol{\theta}^\star$ as $\mathbf{x} \to \mathbf{x}^\star$. ■

C.3 Proof of Proposition 10

Statement. Assume that D_λ given in Theorem 1 is twice differentiable. Let $\mathcal{S} \in \mathbb{R}^m$ be a convex set and $\mathcal{I} = \{i \in [m] : \forall(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \mathcal{S}^2, \theta_i = \theta'_i\}$ a (potentially empty) set of coordinates in which $\mathcal{S}$ reduces to a singleton. Then, D_λ is $\alpha_\mathcal{S}$-strongly concave on $\mathcal{S}$ if and only if

$$0 < \alpha_\mathcal{S} \leq \min_{i\in\mathcal{I}^\complement} -\sup_{\boldsymbol{\theta}\in\mathcal{S}} \sigma_i(\theta_i), \tag{61}$$

where $\sigma_i(\theta_i) = -\lambda^2(f_i^*)''(\lambda\theta_i)$ is the (negative) i-th eigenvalue of the Hessian matrix $\nabla^2 D_\lambda(\boldsymbol{\theta})$.

Proof First of all, let us recall that, from Theorem 1, $D_\lambda(\boldsymbol{\theta}) = -\sum_i f_i^*(-\lambda\theta_i)$. Then by definition of strong concavity, D_λ is α-strongly concave on a set $\mathcal{S} \subset \mathbb{R}^m$ if and only if, $\forall(\boldsymbol{\theta}', \boldsymbol{\theta}) \in \mathcal{S}^2$,

$$D_\lambda(\boldsymbol{\theta}) \leq D_\lambda(\boldsymbol{\theta}') + \langle \nabla D_\lambda(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle - \frac{\alpha^2}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 \tag{62}$$

$$\iff D_\lambda|_{\mathcal{I}^\complement}(\boldsymbol{\theta}_{\mathcal{I}^\complement}) \leq D_\lambda|_{\mathcal{I}^\complement}(\boldsymbol{\theta}'_{\mathcal{I}^\complement}) + \langle \nabla D_\lambda|_{\mathcal{I}^\complement}(\boldsymbol{\theta}'_{\mathcal{I}^\complement}), \boldsymbol{\theta}_{\mathcal{I}^\complement} - \boldsymbol{\theta}'_{\mathcal{I}^\complement} \rangle - \frac{\alpha^2}{2}\|\boldsymbol{\theta}_{\mathcal{I}^\complement} - \boldsymbol{\theta}'_{\mathcal{I}^\complement}\|_2^2 \tag{63}$$

where, for $\mathbf{u} \in \mathbb{R}^{|\mathcal{I}^\complement|}$, $D_\lambda|_{\mathcal{I}^\complement}(\mathbf{u}) = -\sum_{i\in\mathcal{I}^\complement} f_i^*(-\lambda u_i)$ is the restriction of D_λ to the coordinates that belong to $\mathcal{I}^\complement$. The second inequality has been obtained from the definition of $\mathcal{I}$ which implies that $\forall(\boldsymbol{\theta}', \boldsymbol{\theta}) \in \mathcal{S}^2$, $\theta_i = \theta'_i$ for all $i \in \mathcal{I}$. Hence, we have that D_λ is α-strongly concave on $\mathcal{S}$ if and only if its restriction $D_\lambda|_{\mathcal{I}^\complement}$ is α-strongly concave on $\mathcal{S}|_{\mathcal{I}^\complement} = \{\mathbf{u} \in \mathbb{R}^{|\mathcal{I}^\complement|} : \exists \boldsymbol{\theta} \in \mathcal{S} \text{ s.t. } \boldsymbol{\theta}_{\mathcal{I}^\complement} = \mathbf{u}\}$. From Hiriart-Urruty and Lemaréchal (1993a, Chapter IV, Theorem 4.3.1 and Remark 4.3.2), as $\text{int}(\mathcal{S}|_{\mathcal{I}^\complement}) \neq \emptyset$, the latter is equivalent to

$$\alpha \leq \min_{i\in\{1,\dots,|\mathcal{I}^\complement|\}} - \sup_{\mathbf{u}\in\mathcal{S}|_{\mathcal{I}^\complement}} \sigma_i(\nabla^2 D_\lambda|_{\mathcal{I}^\complement}(\mathbf{u})) \tag{64}$$

where $\sigma_i(\nabla^2 D_\lambda|_{\mathcal{I}^\complement}(\mathbf{u}))$ denotes the i-th eigenvalue of the Hessian matrix $(\nabla^2 D_\lambda|_{\mathcal{I}^\complement}(\mathbf{u})) \in \mathbb{R}^{|\mathcal{I}^\complement|\times|\mathcal{I}^\complement|}$. By definition of $D_\lambda|_{\mathcal{I}^\complement}$, we have that for $\boldsymbol{\theta} \in \mathcal{S}$, $\nabla^2 D_\lambda|_{\mathcal{I}^\complement}(\boldsymbol{\theta}_{\mathcal{I}^\complement}) = [\nabla^2 D_\lambda]_{\mathcal{I}^\complement,\mathcal{I}^\complement}(\boldsymbol{\theta})$ where

$$\nabla^2 D_\lambda(\boldsymbol{\theta}) = -\lambda^2 \text{Diag}\left((f_1^*)''(\lambda\theta_1), \dots, (f_m^*)''(\lambda\theta_m)\right). \tag{65}$$

Then, it follows that

$$\alpha \leq \min_{i\in\mathcal{I}^\complement} - \sup_{\boldsymbol{\theta}\in\mathcal{S}} \sigma_i(\theta_i), \tag{66}$$

where $\sigma_i(\theta_i) = -\lambda^2(f_i^*)''(\lambda\theta_i)$. This completes the proof. ■

Appendix D. Particular Cases of Section 4: Calculation Details

In this appendix, we provide details on the derivation of the quantities that are summarized in Table 1.

D.1 Quadratic Distance (β-Divergence with $\beta = 2$)

This corresponds to the standard Lasso problem (when combined with an ℓ_1-norm regularization) and its group-variants. The data-fidelity term in this case is:

$$F(\mathbf{Ax}) = \frac{1}{2}\|\mathbf{y} - \mathbf{Ax}\|_2^2 \qquad f_i([\mathbf{Ax}]_i) = \frac{1}{2}(y_i - [\mathbf{Ax}]_i)^2 \tag{67}$$

with gradient given by

$$\nabla F(\mathbf{Ax}) = \mathbf{Ax} - \mathbf{y} \qquad f_i'([\mathbf{Ax}]_i) = [\mathbf{Ax}]_i - y_i. \tag{68}$$

We then deduce from Theorem 1 that the first-order optimality condition (7) (primal-dual link) is given by:

$$\lambda\boldsymbol{\theta}^\star = \mathbf{y} - \mathbf{A}\mathbf{x}^\star \qquad\qquad \lambda\theta_i^\star = y_i - [\mathbf{A}\mathbf{x}]_i^\star, \quad \forall i \in [m] \tag{69}$$

The maximum regularization parameter $\lambda_{\max}$ is obtained by subtituting $\nabla F(\mathbf{0}) = -\mathbf{y}$ in (20):

$$\lambda_{\max} = \|\mathbf{A}^\mathsf{T}\mathbf{y}\|_\infty \tag{70}$$

The dual function $D_\lambda(\boldsymbol{\theta}) = -\sum_{i=1}^m f_i^*(-\lambda\theta_i)$ is given by:

$$D_\lambda(\boldsymbol{\theta}) = \frac{\|\mathbf{y}\|_2^2 - \|\mathbf{y} - \lambda\boldsymbol{\theta}\|_2^2}{2} = \sum_{i=1}^m \frac{y_i^2 - (y_i - \lambda\theta_i)^2}{2}. \tag{71}$$

with $\mathrm{dom}(D_\lambda) = \mathbb{R}^m$. Then, we get from Theorem 1, $\mathcal{C} = \mathbb{R}^n$, and (19) (dual norm of the ℓ_1-norm) that the dual feasible set is given by:

$$\Delta_{\mathbf{A}} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \|\mathbf{A}^\mathsf{T}\boldsymbol{\theta}\|_\infty \leq 1\}. \tag{72}$$

The Hessian $\nabla^2 D_\lambda(\boldsymbol{\theta})$ and corresponding eigenvalues $\sigma_i(\theta_i)$ are given by

$$\nabla^2 D_\lambda(\boldsymbol{\theta}) = \mathrm{Diag}\left([\sigma_i(\theta_i)]_{i\in[m]}\right), \qquad\qquad \sigma_i\left(\theta_i\right) = -\lambda^2 \tag{73}$$

Hence, from Proposition 10 we get that D_λ is strongly-concave on $\mathbb{R}^m$ with global constant $\alpha = \lambda^2$.

The fact that the dual function is quadratic implies that its second derivative is a constant. Hence, the local strong concavity bound on any subset of $\mathbb{R}^m$ is also λ^2. The proposed local approach thus reduces to the standard Gap Safe screening in this particular case.

Dual Update. The residual w.r.t. a primal estimate $\mathbf{x}$ is given by $\boldsymbol{\rho}(\mathbf{x}) = -\nabla F(\mathbf{A}\mathbf{x}) = \mathbf{y} - \mathbf{A}\mathbf{x}$. Because $\mathcal{S}_0 = \mathbb{R}^m$, given any primal feasible point $\mathbf{x} \in \mathbb{R}^n$ $(= \mathcal{C} \cap \mathrm{dom}(P_\lambda))$, we get from Lemma 9 that

$$\boldsymbol{\Theta}(\mathbf{x}) = \boldsymbol{\Xi}\left((\mathbf{y} - \mathbf{A}\mathbf{x})/\lambda\right). \tag{74}$$

is such that $\boldsymbol{\Theta}(\mathbf{x}) \in \Delta_{\mathbf{A}}$. Moreover, $\boldsymbol{\Theta}(\mathbf{x}) \to \boldsymbol{\theta}^\star$ as $\mathbf{x} \to \mathbf{x}^\star$.

D.2 Logistic Regression

We consider the two-class logistic regression as formulated in Bühlmann and van de Geer (2011, Chapter 3). The observation vector entries $y_i \in \{0, 1\}$ are binary class labels.

The data-fidelity term in this case is:

$$F(\mathbf{A}\mathbf{x}) = \mathbf{1}^\mathsf{T}\log\left(1 + \mathrm{e}^{\mathbf{A}\mathbf{x}}\right) - \mathbf{y}^\mathsf{T}\mathbf{A}\mathbf{x} \qquad f_i([\mathbf{A}\mathbf{x}]_i) = \log\left(1 + \mathrm{e}^{[\mathbf{A}\mathbf{x}]_i}\right) - y_i[\mathbf{A}\mathbf{x}]_i \tag{75}$$

and $\mathcal{C} = \mathbb{R}^n$. The gradient of F is given by

$$\nabla F(\mathbf{z}) = \frac{\mathrm{e}^{\mathbf{z}}}{1 + \mathrm{e}^{\mathbf{z}}} - \mathbf{y} \qquad\qquad f_i'(z_i) = \frac{\mathrm{e}^{z_i}}{1 + \mathrm{e}^{z_i}} - y_i \tag{76}$$

We then deduce from Theorem 1 that the first-order optimality condition (7) (primal-dual link) is given by:

$$\lambda\boldsymbol{\theta}^\star = \mathbf{y} - \frac{\mathrm{e}^{\mathbf{A}\mathbf{x}^\star}}{1+\mathrm{e}^{\mathbf{A}\mathbf{x}^\star}} \qquad \lambda\theta_i^\star = y_i - \frac{\mathrm{e}^{[\mathbf{A}\mathbf{x}^\star]_i}}{1+\mathrm{e}^{[\mathbf{A}\mathbf{x}^\star]_i}}, \quad \forall i \in [m] \tag{77}$$

The maximum regularization parameter $\lambda_{\max}$ is obtained by subtituting $\nabla F(\mathbf{0}) = \frac{1}{2} - \mathbf{y}$ in (20):

$$\lambda_{\max} = \left\| \mathbf{A}^\mathsf{T} \left(\mathbf{y} - 1/2 \right) \right\|_\infty \tag{78}$$

Proposition 20. *(Ndiaye et al., 2017) The Fenchel conjugate of the logistic regression data term in equation* (75) *is given by* $F^*(\mathbf{u}) = \sum_{i=1}^m f_i^*(u_i)$ *where*

$$f_i^*(u) = (y_i + u)\log(y_i + u) + (1 - y_i - u)\log(1 - y_i - u) \tag{79}$$

with $\mathrm{dom}(f_i^*) = [-y_i, 1 - y_i]$.

The dual function $D_\lambda(\boldsymbol{\theta}) = -\sum_{i=1}^m f_i^*(-\lambda\theta_i)$ is given by

$$D_\lambda(\boldsymbol{\theta}) = -(\mathbf{y} - \lambda\boldsymbol{\theta})^\mathsf{T}\log(\mathbf{y} - \lambda\boldsymbol{\theta}) - (1 - \mathbf{y} + \lambda\boldsymbol{\theta})^\mathsf{T}\log(1 - \mathbf{y} + \lambda\boldsymbol{\theta}) \tag{80}$$

with $\mathrm{dom}(D_\lambda) = \{\boldsymbol{\theta} \mid \mathbf{y} - 1 \leq \lambda\boldsymbol{\theta} \leq \mathbf{y}\}$. It is also known as the (binary) entropy function. Then, we get from Theorem 1, $\mathcal{C} = \mathbb{R}^n$, and (19) (dual norm of the ℓ_1-norm) that

$$\Delta_\mathbf{A} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \|\mathbf{A}^\mathsf{T}\boldsymbol{\theta}\|_\infty \leq 1, \mathbf{y} - 1 \leq \lambda\boldsymbol{\theta} \leq \mathbf{y}\} \tag{81}$$

The Hessian $\nabla^2 D_\lambda(\boldsymbol{\theta})$ and corresponding eigenvalues $\sigma_i(\theta_i)$ are given by

$$\nabla^2 D_\lambda(\boldsymbol{\theta}) = \mathrm{Diag}\left([\sigma_i(\theta_i)]_{i\in[m]} \right), \qquad \sigma_i(\theta_i) = -\frac{\lambda^2}{(y_i - \lambda\theta_i)(1 - y_i + \lambda\theta_i)} \tag{82}$$

$$= -\frac{4\lambda^2}{1 - 4(\lambda\theta_i - y_i + \frac{1}{2})^2} \tag{83}$$

The eigenvalues $\sigma_i(\theta_i)$ are all negative with maximum value $-4\lambda^2$ attained at $\lambda\theta_i = (2y_i - 1)/2$. Therefore, the dual function is strongly concave on $\mathrm{dom}(D_\lambda)$ with a global constant

$$\alpha = 4\lambda^2.$$

Let us now examine the dual function restricted to some particular subsets S of the domain $\mathrm{dom}(D_\lambda)$.

D.2.1 Strong-Concavity Bound on $\Delta_A \cap \mathcal{S}_0$

In this section we evaluate the strong concavity of D_λ on the set $S = \Delta_\mathbf{A}$ or equivalently $S = \Delta_\mathbf{A} \cap \mathcal{S}_0$ with $\mathcal{S}_0 = \mathbb{R}^m$.

Proposition 21. *Assuming that* $\text{rank}(\mathbf{A}) = \min(m, n)$*, the dual function* D_λ *as defined in* (80) *is* $\alpha_{\Delta_\mathbf{A}}$ *strongly concave on* $\Delta_\mathbf{A}$ *with constant:*

$$\alpha_{\Delta_\mathbf{A}} = \frac{4\lambda^2}{1 - 4\left(\min(\lambda\|\mathbf{A}^\dagger\|_1, \frac{1}{2}) - \frac{1}{2}\right)^2} \tag{84}$$

where $\mathbf{A}^\dagger$ *denotes the right pseudo-inverse of* $\mathbf{A}$ *such that* $\mathbf{A}\mathbf{A}^\dagger = \mathbf{I}$*, and* $\|\mathbf{A}^\dagger\|_1$ *is the maximum absolute column sum of* $\mathbf{A}^\dagger$*.*
Moreover, the local bound $\alpha_{\Delta_\mathbf{A}}$ *in* (84) *improves upon the global bound* $\alpha = 4\lambda^2$ *for all* $\lambda < \frac{1}{2\|\mathbf{A}^\dagger\|_1}$*.*

Proof From Proposition 10 (with $\mathcal{I} = \emptyset$) we have to prove that

$$\alpha_{\Delta_\mathbf{A}} \leq \min_{i\in[m]} - \sup_{\boldsymbol{\theta}\in\Delta_A} \sigma_i(\theta_i). \tag{85}$$

where $\sigma_i(\theta_i) = -\frac{\lambda^2}{(y_i - \lambda\theta_i)(1 - y_i + \lambda\theta_i)}$ is the i-th eigenvalue of $\nabla^2 D_\lambda(\boldsymbol{\theta})$. Given that $y_i \in \{0, 1\}$, we can compactly rewrite $\sigma_i(\theta_i)$ as follows:

$$\sigma_i(\theta_i) = \begin{cases} -\frac{\lambda^2}{(1-\lambda\theta_i)(\lambda\theta_i)} & \forall \lambda\theta_i \in (0,1) \quad \text{if } y_i = 1 \\ -\frac{\lambda^2}{(-\lambda\theta_i)(1+\lambda\theta_i)} & \forall \lambda\theta_i \in (-1,0) \quad \text{if } y_i = 0 \end{cases} \tag{86}$$

$$= -\frac{\lambda^2}{\lambda|\theta_i|(1 - \lambda|\theta_i|)} \quad \forall |\lambda\theta_i| \in (0,1) \tag{87}$$

First, note that if $\lambda\|\boldsymbol{\theta}\|_\infty \geq 1/2$ then the global maximum of $\sigma_i(\theta_i)$ is attained for some i. Indeed, the global maximum is attained for $\lambda|\theta_i| = 1/2$:

$$\lambda\theta_i = y_i - \frac{1}{2} = \begin{cases} 1/2 & \text{if } y_i = 1 \\ -1/2 & \text{if } y_i = 0 \end{cases}$$

For $\lambda|\theta_i| \leq 1/2$, one can see that $\sigma_i(\theta_i)$ is increasing with $|\theta_i|$, which implies that

$$\min_{i\in[m]} - \sup_{\|\boldsymbol{\theta}\|_\infty \leq a} \sigma_i(\theta_i) = \frac{\lambda^2}{\min(\lambda a, \frac{1}{2})(1 - \min(\lambda a, \frac{1}{2}))} = \begin{cases} \frac{\lambda}{a(1-\lambda a)} & \text{if } \lambda a < \frac{1}{2} \\ 4\lambda^2 & \text{otherwise} \end{cases} \tag{88}$$

Now, we can bound $\|\boldsymbol{\theta}\|_\infty$ on $\Delta_\mathbf{A}$ as follows, knowing that $\|\mathbf{A}^\mathsf{T}\boldsymbol{\theta}\|_\infty \leq 1$ and $\text{rank}(\mathbf{A}) = \min(m, n)$:

$$\begin{aligned} \|\boldsymbol{\theta}\|_\infty = \max_i |\theta_i| = \max_i |\left[(\mathbf{A}\mathbf{A}^\dagger)^\mathsf{T}\boldsymbol{\theta}\right]_i| = \max_i |\langle (\mathbf{A}^\dagger)_i, \mathbf{A}^\mathsf{T}\boldsymbol{\theta}\rangle| \\ \leq \max_i \|(\mathbf{A}^\dagger)_i\|_1 \|\mathbf{A}^\mathsf{T}\boldsymbol{\theta}\|_\infty = \|\mathbf{A}^\dagger\|_1 \|\mathbf{A}^\mathsf{T}\boldsymbol{\theta}\|_\infty \leq \|\mathbf{A}^\dagger\|_1 \end{aligned}$$

Hence, we have that $\Delta_\mathbf{A} \subseteq \{\boldsymbol{\theta} \in \mathbb{R}^m | \|\boldsymbol{\theta}\|_\infty \leq \|\mathbf{A}^\dagger\|_1\}$ and

$$\begin{aligned} \min_{i\in[m]} - \sup_{\boldsymbol{\theta}\in\Delta_A} \sigma_i(\theta_i) \geq \min_{i\in[m]} - \sup_{\|\boldsymbol{\theta}\|_\infty \leq \|\mathbf{A}^\dagger\|_1} \sigma_i(\theta_i) &= \frac{\lambda^2}{\min(\lambda\|\mathbf{A}^\dagger\|_1, \frac{1}{2})(1 - \min(\lambda\|\mathbf{A}^\dagger\|_1, \frac{1}{2}))} \\ &= \frac{4\lambda^2}{1 - 4\left(\min(\lambda\|\mathbf{A}^\dagger\|_1, \frac{1}{2}) - \frac{1}{2}\right)^2} \\ &= \alpha_{\Delta_\mathbf{A}}, \end{aligned}$$

where the first equality is obtained by taking $a = \|\mathbf{A}^\dagger\|_1$ in the equation (88) Finally, one can easily verify that $\alpha_{\Delta_\mathbf{A}} < 4\lambda^2 \iff \lambda < \frac{1}{2\|\mathbf{A}^\dagger\|_1}$. ■

D.2.2 Strong-Concavity Bound on $\mathcal{B}(\theta, r) \cap \mathcal{S}_0$

Proposition 22. *For any ball $\mathcal{B}(\boldsymbol{\theta}, r)$ with $\boldsymbol{\theta} \in \mathrm{dom}(D_\lambda)$, the dual function D_λ as defined in (80) is $\alpha_{\mathcal{B}(\boldsymbol{\theta},r)}$ strongly concave on $\mathcal{B}(\boldsymbol{\theta}, r)$ with constant:*

$$\alpha_{\mathcal{B}(\boldsymbol{\theta},r)} = \frac{4\lambda^2}{1 - 4([\min_i(|\lambda\theta_i - y_i + \frac{1}{2}|) - \lambda r]^+)^2} \tag{89}$$

Proof From Proposition 10 (with $\mathcal{I} = \emptyset$) we have to prove that

$$\alpha_{\mathcal{B}(\boldsymbol{\theta},r)} \leq \min_{i\in[m]} - \sup_{\boldsymbol{\theta}' \in \mathcal{B}(\boldsymbol{\theta},r)} \sigma_i(\theta'_i). \tag{90}$$

where the i-th eigenvalue of $\nabla^2 D_\lambda(\boldsymbol{\theta})$ is given by

$$\sigma_i(\theta_i) = -\frac{\lambda^2}{(y_i - \lambda\theta_i)(1 - y_i + \lambda\theta_i)} = -\frac{4\lambda^2}{1 - 4(\lambda\theta_i - y_i + \frac{1}{2})^2}, \quad \lambda\theta_i \in (y_i - 1, y_i)$$

The maximum value $\sup \sigma_i(\theta_i) = 4\lambda^2$ attained at $\lambda\theta_i = y_i - \frac{1}{2} := a^\star$. Also note that that $\sigma_i(\theta_i)$ symmetric w.r.t. $a^\star/\lambda$ and a decreasing function w.r.t. $|\theta_i - a^\star/\lambda|$ (i.e. $\sigma_i(\theta_i)$ only decreases as $\lambda\theta_i$ gets further away from $a^\star$).

Now we evaluate $\sup_{\boldsymbol{\theta}' \in \mathcal{B}(\boldsymbol{\theta},r)} \sigma_i(\theta'_i) = \sup_{|\theta_i - \theta'_i| \leq r} \sigma_i(\theta'_i)$ for the i-th eigenvalue. If $|\theta_i - a^\star/\lambda| < r$, the global maximum is attained as $a^\star/\lambda$ lies inside the interval. Otherwise, the maximum lies on the border of the interval which is closest to $a^\star/\lambda$, i.e. $\theta'_i = \theta_i - r$ if $\theta_i > a^\star/\lambda + r$ and $\theta'_i = \theta_i + r$ if $\theta_i < a^\star/\lambda - r$.

$$\begin{aligned}
\sup_{|\theta_i - \theta'_i| \leq r} \sigma_i(\theta'_i) &= \begin{cases} -4\lambda^2 & \text{if } -r < \theta_i - a^\star/\lambda < r \\ -\frac{4\lambda^2}{1-4(\lambda(\theta_i - r) - a^\star)^2} & \text{if } \theta_i - a^\star/\lambda \geq r \\ -\frac{4\lambda^2}{1-4(\lambda(\theta_i + r) - a^\star)^2} & \text{if } \theta_i - a^\star/\lambda \leq -r \end{cases} \\
&= \begin{cases} -4\lambda^2 & \text{if } |\lambda\theta_i - a^\star| < \lambda r \\ -\frac{4\lambda^2}{1-4(|\lambda\theta_i - a^\star| - \lambda r)^2} & \text{if } |\lambda\theta_i - a^\star| \geq \lambda r \end{cases}
\end{aligned}$$

Finally, $\alpha_{B(\boldsymbol{\theta},r)}$ is obtained by

$$\begin{aligned}
\alpha_{B(\boldsymbol{\theta},r)} &\leq \min_{i\in[m]} - \sup_{|\theta_i - \theta'_i| \leq r} \sigma_i(\theta'_i) \\
&= \begin{cases} 4\lambda^2 & \text{if } \min_i(|\lambda\theta_i - a^\star|) < \lambda r \\ \frac{4\lambda^2}{1-4(\min_i |\lambda\theta_i - a^\star| - \lambda r)^2} & \text{if } \min_i(|\lambda\theta_i - a^\star|) \geq \lambda r \end{cases} \\
&= \begin{cases} 4\lambda^2 & \text{if } \min_i(|\lambda\theta_i - y_i + \frac{1}{2}|) \leq \lambda r \\ \frac{4\lambda^2}{1-4(\min_i(|\lambda\theta_i - y_i + \frac{1}{2}|) - \lambda r)^2} & \text{otherwise.} \end{cases} \\
&= \frac{4\lambda^2}{1 - 4([\min_i(|\lambda\theta_i - y_i + \frac{1}{2}|) - \lambda r]^+)^2}
\end{aligned}$$

This completes the proof. ■

D.2.3 Dual Update

The generalized residual w.r.t. a primal estimate $\mathbf{x}$ is given by

$$\boldsymbol{\rho}(\mathbf{x}) := -\nabla F(\mathbf{A}\mathbf{x}) = \mathbf{y} - \frac{\mathrm{e}^{\mathbf{A}\mathbf{x}}}{1+\mathrm{e}^{\mathbf{A}\mathbf{x}}}.$$

Because $\mathcal{S}_0 = \mathbb{R}^m$, given any primal feasible point $\mathbf{x} \in \mathcal{C} \cap \mathrm{dom}(P_\lambda)$, we get from Lemma 9 that

$$\boldsymbol{\Theta}(\mathbf{x}) = \boldsymbol{\Xi}\left(\frac{1}{\lambda}\left(\mathbf{y} - \frac{\mathrm{e}^{\mathbf{A}\mathbf{x}}}{1+\mathrm{e}^{\mathbf{A}\mathbf{x}}}\right)\right) \tag{91}$$

is such that $\boldsymbol{\Theta}(\mathbf{x}) \in \Delta_{\mathbf{A}}$. Moreover, $\boldsymbol{\Theta}(\mathbf{x}) \to \boldsymbol{\theta}^\star$ as $\mathbf{x} \to \mathbf{x}^\star$.

D.3 β-Divergence with $\beta \in (1,2)$

The data fidelity term is given by the β-divergence between the input signal $\mathbf{y} \in \mathbb{R}^m_+$ and its reconstruction $\mathbf{A}\mathbf{x}$, i.e.

$$F(\mathbf{A}\mathbf{x}) = \frac{1}{\beta(\beta-1)}\left(\|\mathbf{y}\|_\beta^\beta + (\beta-1)\|\mathbf{A}\mathbf{x}+\epsilon\|_\beta^\beta - \beta\mathbf{y}^\mathsf{T}(\mathbf{A}\mathbf{x}+\epsilon)^{\beta-1}\right) \tag{92}$$

$$f_i([\mathbf{A}\mathbf{x}]_i) = \frac{1}{\beta(\beta-1)}\left(y_i^\beta + (\beta-1)([\mathbf{A}\mathbf{x}]_i+\epsilon)^\beta - \beta y_i([\mathbf{A}\mathbf{x}]_i+\epsilon)^{\beta-1}\right) \tag{93}$$

Note that we introduce an ϵ-smoothing factor ($\epsilon > 0$) on the second variable. This is a common practice in the literature (Harmany et al., 2012) to avoid singularities around zero.

For the above equations to be well-defined we need that $\mathbf{A}\mathbf{x} + \epsilon \geq 0$ for all $\mathbf{x} \in \mathcal{C}$. We therefore take $\mathcal{C} = \mathbb{R}^n_+$. Moreover, we consider that $\mathbf{A} \in \mathbb{R}^{m\times n}_+$.

The first derivative is given by:

$$\nabla F(\mathbf{z}) = (\mathbf{z}+\epsilon)^{\beta-2}(\mathbf{z}+\epsilon-\mathbf{y}) \qquad f_i'(z_i) = (z_i+\epsilon)^{\beta-2}(z_i+\epsilon-y_i) \tag{94}$$

We then deduce from Theorem 1 that the first-order optimality condition (7) (primal-dual link) is given by:

$$\lambda\theta_i^\star = ([\mathbf{A}\mathbf{x}^\star]_i+\epsilon)^{\beta-2}(y_i - [\mathbf{A}\mathbf{x}^\star]_i - \epsilon), \quad \forall i \in [m] \tag{95}$$

The maximum regularization parameter $\lambda_{\max}$ is obtained by subtituting $\nabla F(\mathbf{0}) = \epsilon^{\beta-2}(\epsilon-\mathbf{y})$ in (20):

$$\lambda_{\max} = \epsilon^{\beta-2}\max(\mathbf{A}^\mathsf{T}(\mathbf{y}-\epsilon)) \tag{96}$$

Proposition 23. *The Fenchel conjugate of the β-divergence data-fidelity function defined in (92) is given by $F^*(\mathbf{u}) = \sum_{i=1}^m f_i^*(u_i)$ where*

$$f_i^*(u_i) = \begin{cases} -\epsilon u_i, & \text{if } \ y_i = 0, u_i \leq 0 \\ u_i\hat{z}_i - f(\hat{z}_i), & \text{otherwise} \end{cases} \tag{97}$$

with $\hat{z}_i \geq 0$ such that

$$(\hat{z}_i+\epsilon)^{\beta-2}(\hat{z}_i+\epsilon-y_i) = u_i \tag{98}$$

Proof Given that $\text{dom}(f_i) = \{z_i \in \mathbb{R} \mid z_i + \epsilon \geq 0\}$, the Fenchel conjugate of f_i is given by

$$f_i^*(u_i) = \sup_{z_i \in \mathbb{R}} \underbrace{z_i u_i - f_i(z_i)}_{\varphi(z_i)} = \sup_{z_i + \epsilon \geq 0} \varphi(z_i). \tag{99}$$

Because f_i is a convex function, we have that φ is concave. Hence, every stationary point $\hat{z}_i$ such that $\hat{z}_i + \epsilon \geq 0$ and $0 = \varphi'(\hat{z}_i) = u_i - (\hat{z}_i + \epsilon)^{\beta-2}(\hat{z}_i + \epsilon - y_i)$ is a global maximum. We now distinguish two cases:

- When $y_i > 0$, $\varphi'(z_i) \to +\infty$ as $z_i \to -\epsilon$ and $\varphi'(z_i) \to -\infty$ as $z_i \to +\infty$, which combined to the fact that φ' is continuous and decreasing (as φ is concave) implies that there exists $\hat{z}_i \geq -\epsilon$ such that $\varphi'(\hat{z}_i) = 0$.
- When $y_i = 0$, we have $\varphi(z_i) = z_i u_i - \frac{1}{\beta}(z_i + \epsilon)^\beta$ and $\varphi'(\hat{z}_i) = 0 \Rightarrow \hat{z}_i + \epsilon = u_i^{1-\beta}$. It follows that $\hat{z}_i$ satisfies the constraints in (99) if and only if $u_i \geq 0$. In contrast, when $u_i < 0$ the sup in (99) is attained at $z_i + \epsilon = 0$ with value $-\epsilon u_i$.

This completes the proof. ■

Unfortunately, equation (98) does not admit an explicit closed-form expression without fixing a value for β. This prevents the derivation of an explicit form of f_i^* in function of β. We thus focus in the case $\beta = 1.5$ in the remainder of this section.

D.3.1 Particular Case of $\beta = 1.5$

Taking $\beta = 1.5$ on the previous results we obtain:

$$F(\mathbf{Ax}) = \frac{4}{3}\left(\|\mathbf{y}\|_{1.5}^{1.5} + \frac{1}{2}\|\mathbf{Ax} + \epsilon\|_{1.5}^{1.5} - \frac{3}{2}\mathbf{y}^\mathsf{T}(\mathbf{Ax} + \epsilon)^{0.5}\right) \tag{100}$$

$$f_i([\mathbf{Ax}]_i) = \frac{4}{3}\left(y_i^{1.5} + \frac{1}{2}([\mathbf{Ax}]_i + \epsilon)^{1.5} - \frac{3}{2}y_i([\mathbf{Ax}]_i + \epsilon)^{0.5}\right) \tag{101}$$

with first derivative is given by

$$\nabla F(\mathbf{z}) = \sqrt{\mathbf{z} + \epsilon} - \frac{\mathbf{y}}{\sqrt{\mathbf{z} + \epsilon}} \qquad f_i'(z_i) = \sqrt{z_i + \epsilon} - \frac{y_i}{\sqrt{z_i + \epsilon}}. \tag{102}$$

The first optimality condition becomes:

$$\lambda\theta_i^\star = \frac{y_i}{\sqrt{[\mathbf{Ax}^\star]_i + \epsilon}} - \sqrt{[\mathbf{Ax}^\star]_i + \epsilon}, \quad \forall i \in [m] \tag{103}$$

The maximum regularization parameter $\lambda_{\max}$ is obtained by subtituting $\nabla F(\mathbf{0}) = \frac{\epsilon - \mathbf{y}}{\sqrt{\epsilon}}$ in (20):

$$\lambda_{\max} = \max(\mathbf{A}^\mathsf{T}(\mathbf{y} - \epsilon)/\sqrt{\epsilon}) \tag{104}$$

Proposition 24. *The Fenchel conjugate of the β-divergence with $\beta = 1.5$ in equation* (100) *is given by $F^*(\mathbf{u}) = \sum_{i=1}^m f_i^*(u_i)$ where*

$$f_i^*(u_i) = \frac{u_i^3}{6} + \frac{1}{6}(u_i^2 + 4y_i)^{1.5} + u_i y_i - \frac{4}{3}y_i^{1.5} - \epsilon u_i \tag{105}$$

with $\text{dom}(f_i^*) = \mathbb{R}$.

Proof According to Proposition 23, the fenchel conjugate for $\beta = 1.5$ is given by

$$f_i^*(u_i) = \begin{cases} -\epsilon u_i, & \text{if } \; y_i = 0, u_i \leq 0 \\ u_i \hat{z}_i - f(\hat{z}_i), & \text{otherwise} \end{cases}$$

where $\hat{z}_i$ is the solution of the following equation:

$$\frac{(\hat{z}_i + \epsilon - y_i)}{\sqrt{\hat{z}_i + \epsilon}} = u_i. \tag{106}$$

It gives a quadratic equation on $\sqrt{\hat{z}_i + \epsilon}$ with solution

$$\sqrt{\hat{z}_i + \epsilon} = \frac{u_i \pm \sqrt{u_i^2 + 4y_i}}{2} \implies \hat{z}_i = \left(\frac{u_i + \sqrt{u_i^2 + 4y_i}}{2} \right)^2 - \epsilon \tag{107}$$

since $u_i \leq \sqrt{u_i^2 + 4y_i}$ for $y_i \geq 0$. Plugging $\hat{z}_i$ in (106) we obtain after some calculation:

$$f_i^*(u_i) = \begin{cases} -\epsilon u_i, & \text{if } \; y_i = 0, u_i \leq 0 \\ \frac{u_i^3}{6} + \frac{1}{6}(u_i^2 + 4y_i)^{1.5} + u_i y_i - \frac{4}{3} y_i^{1.5} - \epsilon u_i, & \text{otherwise} \end{cases} \tag{108}$$

Finally, note that taking $y_i = 0$ and $u_i \leq 0$ in (105) we obtain $f_i^*(u_i) = -\epsilon u_i$, in accordance with equation (108). Therefore, (105) alone summarizes both cases depicted in (108). ■

The dual function $D_\lambda(\boldsymbol{\theta}) = \sum_{i=1}^m -f_i^*(-\lambda\theta_i)$ is given by:

$$D_\lambda(\boldsymbol{\theta}) = \sum_{i=1}^m \frac{1}{6}(\lambda\theta_i)^3 - \frac{1}{6}\left((\lambda\theta_i)^2 + 4y_i\right)^{1.5} + \lambda\theta_i y_i + \frac{4}{3} y_i^{1.5} - \epsilon\lambda\theta_i \tag{109}$$

with $\mathrm{dom}(D_\lambda) = \mathbb{R}^m$. Then, we get from Theorem 1, $\mathcal{C} = \mathbb{R}^n_+$, and (19) (dual norm of the ℓ_1-norm) that the dual feasible set is given by:

$$\Delta_{\mathbf{A}} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \mathbf{A}^\mathsf{T}\boldsymbol{\theta} \leq \mathbf{1}\} \tag{110}$$

The Hessian $\nabla^2 D_\lambda(\boldsymbol{\theta})$ and corresponding eigenvalues $\sigma_i(\theta_i)$ are given by

$$\nabla^2 D_\lambda(\boldsymbol{\theta}) = \mathrm{Diag}\left([\sigma_i(\theta_i)]_{i\in[m]}\right), \qquad \sigma_i(\theta_i) = -\lambda^2 \left(\frac{(\lambda\theta_i)^2 + 2y_i}{\sqrt{(\lambda\theta_i)^2 + 4y_i}} - \lambda\theta_i \right) \tag{111}$$

Although the eigenvalues are all non-positive, they tend to zero as θ_i tends to infinity. Note that $\sigma_i(\theta_i)$ also vanishes when $y_i = 0$ and $\theta_i \geq 0$. Therefore $D_\lambda(\boldsymbol{\theta})$ is not globally strongly concave and the standard Gap Safe approach cannot be applied in this case.

In the following sections we find local strong concavity bounds that allows us to deploy the proposed screening strategy (Algorithms 2 and 3).

Proposition 25. *The eigenvalues $\sigma_i(\theta_i)$ in (111) are an increasing (resp. strictly increasing) function of θ_i for any $y_i \geq 0$ (resp. $y_i > 0$).*

Proof Indeed, its first derivative is non-negative

$$\sigma_i'(\theta_i) = \lambda^3 \frac{((\lambda\theta_i)^2 + 4y_i)^{1.5} - (\lambda\theta_i)^3 - 6y_i\lambda\theta_i}{((\lambda\theta_i)^2 + 4y_i)^{1.5}} \geq 0 \tag{112}$$

since (let us recall that $\mathbf{y} \in \mathbb{R}^m_+$) for $\lambda\theta_i \leq 0$ we have $((\lambda\theta_i)^2 + 4y_i)^{1.5} \geq 0 \geq (\lambda\theta_i)^3 + 6y_i\lambda\theta_i$ and for $\lambda\theta_i > 0$ one can easily verify that $\big((\lambda\theta_i)^2 + 4y_i)^{1.5}\big)^2 \geq \big((\lambda\theta_i)^3 + 6y_i\lambda\theta_i\big)^2$ with strict inequalities when $y_i > 0$. ■

D.3.2 Validity of the Set $\mathcal{S}_0$

In order to be able to derive local strong concavity bounds (see Sections D.3.3 and D.3.4 hereafter), we need to use the set $\mathcal{S}_0$ below (see discussion in Section 4.2.1). To comply with Theorem 5, we need to show that $\boldsymbol{\theta}^\star \in \mathcal{S}_0$.

Proposition 26. *For any $\mathbf{A} \in \mathbb{R}^{m\times n}_+$ with no all-zero row, $\mathbf{y} \in \mathbb{R}^m_+$, the set*

$$\mathcal{S}_0 = \big\{\boldsymbol{\theta} \in \mathbb{R}^m \;\big|\; \lambda\boldsymbol{\theta} \leq \min(\mathbf{b}, (\mathbf{y}-\epsilon)/\sqrt{\epsilon})\big\} \tag{113}$$

$$\textit{with}\;\; b_i := \lambda \min_{\{j\in[n] \;|\; a_{ij}\neq 0\}} \left(\frac{1 - c\|\mathbf{a}_j\|_1}{a_{ij}}\right) + \lambda c, \quad i \in [m]$$

$$c := -\frac{1}{\lambda}\sqrt[3]{\frac{4\|\mathbf{y}\|_{1.5}^{1.5} + 2(m-1)\epsilon^{1.5} + 3\epsilon}{1-3\epsilon}}.$$

is such that $\boldsymbol{\theta}^\star \in \mathcal{S}_0$.

Proof Because $\mathbf{A} \in \mathbb{R}^{m\times n}_+$ and $\mathbf{x}^\star \in \mathbb{R}^n_+$, we get from the optimality condition (103) that

$$\lambda\boldsymbol{\theta}^\star = \frac{\mathbf{y} - \mathbf{A}\mathbf{x}^\star - \epsilon}{\sqrt{\mathbf{A}\mathbf{x}^\star + \epsilon}} \leq \frac{\mathbf{y}-\epsilon}{\sqrt{\epsilon}}$$

In particular, for coordinates $\mathcal{I}_0 = \{i \in [m] \mid y_i = 0\}$, this inequality simplifies to $\lambda\boldsymbol{\theta}_{\mathcal{I}_0} \leq -\sqrt{\epsilon}$ (which is strictly negative, as desired to prevent Hessian eigenvalues from vanishing). Let us emphasize that defining $\mathcal{S}_0 = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \lambda\boldsymbol{\theta} \leq (\mathbf{y}-\epsilon)/\sqrt{\epsilon}\}$ is already a valid choice. However, combining it with the tricky bound $\mathbf{b}$ (derived hereafter) can lead to improved strong concavity bounds, which are particularly relevant for Algorithm 2.

Then, to show that $\lambda\theta_i^\star \leq b_i$ we will start by finding a lower bound $c_i \leq \theta_i^\star$ which, combined with the definition of the feasible set, will directly lead to the desired upper bounds b_i. Consider the dual point $\boldsymbol{\theta} = \mathbf{0}$, which is always feasible, then we have:

$$D_\lambda(\boldsymbol{\theta}^\star) \geq D_\lambda(\mathbf{0}) = 0 \tag{114}$$

Also, denoting for simplicity $d_i(\theta_i) = -f_i^*(-\lambda\theta_i)$ such that $D_\lambda(\boldsymbol{\theta}) = \sum_i d_i(\theta_i)$, one can verify that:

$$\sup_{\theta_i} d_i(\theta_i) = \frac{4}{3}y_i^{1.5} - 2y_i\sqrt{\epsilon} + \frac{2}{3}\epsilon^{1.5} \leq \frac{4}{3}y_i^{1.5} + \frac{2}{3}\epsilon^{1.5} \tag{115}$$

Indeed, $d_i(\theta_i)$ being concave we have that the stationary point $\theta_i = \frac{y_i - \epsilon}{\sqrt{\epsilon}}$ (obtained by setting $d'_\lambda(\theta_i) = 0$ with the help of some symbolic calculation tool) is the global maximum, with value $d_i(\theta_i) = \frac{4}{3}y_i^{1.5} - 2y_i\sqrt{\epsilon} + \frac{2}{3}\epsilon^{1.5}$. Combining (114) and (115) we obtain

$$\begin{aligned} d_i(\theta_i^\star) = D_\lambda(\boldsymbol{\theta}^\star) - \sum_{i' \neq i} d_{i'}(\theta_{i'}^\star) &\geq 0 - \sum_{i' \neq i} d_{i'}(\theta_{i'}^\star) \\ &\geq -\sum_{i' \neq i} \left(\frac{4}{3} y_{i'}^{1.5} + \frac{2}{3}\epsilon^{1.5} \right) = -\frac{4}{3}(\|\mathbf{y}\|_{1.5}^{1.5} - y_i^{1.5}) - \frac{2}{3}(m-1)\epsilon^{1.5} \end{aligned}$$

Since $d_i(\theta_i)$ is concave, continuous, and $\lim_{\theta_i \to -\infty} d_i(\theta_i) = -\infty$, it implies that there exists a $\hat{c}_i \leq \theta_i^\star$ such that $d_i(\hat{c}_i) = -\frac{4}{3}(\|\mathbf{y}\|_{1.5}^{1.5} - y_i^{1.5}) - \frac{2}{3}(m-1)\epsilon^{1.5}$ and which can obtained by solving the resulting equation for $\hat{c}_i$.

This, however, leads to a cumbersome calculation. Instead, we provide another bound $c_i \leq \hat{c}_i \leq \theta_i^\star$. To do so, we use an upper bound $g \geq d_i$ and define c_i such that $g(c_i) = d_i(\hat{c}_i)$. We then have $g(c_i) = d_i(\hat{c}_i) \leq g(\hat{c}_i)$ and because we choose g an increasing function, it implies that $c_i \leq \hat{c}_i$ as desired. Noting that $\hat{c}_i \leq 0$ since $-\frac{4}{3}(\|\mathbf{y}\|_{1.5}^{1.5} - y_i^{1.5}) - \frac{2}{3}(m-1)\epsilon^{1.5} \leq 0 = d_i(0)$, we use the following bound on d_i valid for any $\theta_i \leq 0$:

$$\begin{aligned} (\forall\ \theta_i \leq 0)\quad d_i(\theta_i) &= \frac{1}{6}(\lambda\theta_i)^3 - \frac{1}{6}\left((\lambda\theta_i)^2 + 4y_i\right)^{1.5} + \lambda\theta_i y_i + \frac{4}{3}y_i^{1.5} - \epsilon\lambda\theta_i \\ &\leq \frac{1}{6}(\lambda\theta_i)^3 - \frac{1}{6}|\lambda\theta_i|^3 + \lambda\theta_i y_i + \frac{4}{3}y_i^{1.5} - \epsilon\lambda\theta_i \\ &\leq \frac{1}{3}(\lambda\theta_i)^3 + \frac{4}{3}y_i^{1.5} - \epsilon\lambda\theta_i \\ &\leq \left(\frac{1}{3} - \epsilon\right)(\lambda\theta_i)^3 + \frac{4}{3}y_i^{1.5} + \epsilon = g(\theta_i) \end{aligned}$$

where in the last step we used the fact that $-\epsilon\lambda\theta_i \leq -\epsilon\left((\lambda\theta_i)^3 - 1\right)$, $\forall \theta_i \leq 0$. We can now compute c_i such that $g(c_i) = d_i(\hat{c}_i) = -\frac{4}{3}(\|\mathbf{y}\|_{1.5}^{1.5} - y_i^{1.5}) - \frac{2}{3}(m-1)\epsilon^{1.5}$ by solving the following equation for c_i:

$$\left(\frac{1}{3} - \epsilon\right)(\lambda c_i)^3 + \frac{4}{3}y_i^{1.5} + \epsilon = -\frac{4}{3}(\|\mathbf{y}\|_{1.5}^{1.5} - y_i^{1.5}) - \frac{2}{3}(m-1)\epsilon^{1.5}$$

$$\implies c := c_i = -\frac{1}{\lambda}\sqrt[3]{\frac{4\|\mathbf{y}\|_{1.5}^{1.5} + 2(m-1)\epsilon^{1.5} + 3\epsilon}{1 - 3\epsilon}} \leq \theta_i^\star.$$

As c_i turns out to be independent of i, we denote simply c such that $c \leq \theta_i^\star$ for all $i \in [m]$.

Now, we can upper bound $\theta_i^\star$ for any $i \in [m]$ by combining the above lower bound with fact that $\mathbf{A}^\mathsf{T}\boldsymbol{\theta}^\star \leq \mathbf{1}$ (since $\boldsymbol{\theta}^\star \in \Delta_\mathbf{A}$). For all $i \in [m]$ and $\mathbf{A} \in \mathbb{R}_+^{m \times n}$ we have:

$$\forall j \in [n],\quad 1 \geq \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}^\star = \sum_{i'=1}^{m} a_{i'j}\theta_{i'}^\star = a_{ij}\theta_i^\star + \sum_{i' \neq i} a_{i'j}\theta_{i'}^\star \geq a_{ij}\theta_i^\star + \sum_{i' \neq i} a_{i'j}c$$

which leads to the following upper bound b_i for $\lambda\theta_i^\star$, with $i \in [m]$:

$$\forall j \in [n], \quad a_{ij}\theta_i^\star \leq 1 - \sum_{i' \neq i} a_{i'j}c = 1 - c(\|\mathbf{a}_j\|_1 - a_{ij}) \tag{116}$$

$$\iff \quad \theta_i^\star \leq \min_{\{j \in [n] \mid a_{ij} \neq 0\}} \left(\frac{1 - c\|\mathbf{a}_j\|_1}{a_{ij}} \right) + c := b_i/\lambda \tag{117}$$

where our assumption that $\mathbf{A}$ has no all-zero row implies that the set $\{j \in [n] \mid a_{ij} \neq 0\}$ is non-empty. Finally, because $b_i > 0 \ \forall i$, the inequality $\lambda\boldsymbol{\theta}_{\mathcal{I}_0} \leq -\sqrt{\epsilon}$ is always more restrictive for coordinates $i \in \mathcal{I}_0$. ■

D.3.3 Strong-Concavity Bound on $\Delta_A \cap \mathcal{S}_0$

Proposition 27. *Let $\mathcal{S}_0$ be the set defined in Proposition 26 and $\mathbf{A} \in \mathbb{R}_+^{m \times n}$. Then the dual function D_λ as defined in (109) is $\alpha_{\Delta_\mathbf{A} \cap \mathcal{S}_0}$ strongly concave on $\Delta_\mathbf{A} \cap \mathcal{S}_0$ with constant:*

$$\alpha_{\Delta_\mathbf{A} \cap \mathcal{S}_0} = \min_{i \in [m]} -\sigma_i \left(\frac{1}{\lambda} \min \left(b_i, \frac{y_i - \epsilon}{\sqrt{\epsilon}} \right) \right), \tag{118}$$

where the b_i are quantities that define $\mathcal{S}_0$.

Proof From Proposition 10 (with $\mathcal{I} = \emptyset$) we have to prove that

$$\alpha_{\Delta_\mathbf{A} \cap \mathcal{S}_0} \leq \min_{i \in [m]} - \sup_{\boldsymbol{\theta} \in \Delta_\mathbf{A} \cap \mathcal{S}_0} \sigma_i(\theta_i). \tag{119}$$

Here, we have $\Delta_\mathbf{A} \cap \mathcal{S}_0 = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \mathbf{A}^\top\boldsymbol{\theta} \leq \mathbf{1}, \ \lambda\boldsymbol{\theta} \leq \mathbf{b}, \lambda\boldsymbol{\theta} \leq \frac{\mathbf{y} - \epsilon}{\sqrt{\epsilon}}\}$.

Within $\mathcal{S}_0$ the coordinates θ_i are upper bounded with $\lambda\theta_i \leq \min\left(b_i, \frac{y_i - \epsilon}{\sqrt{\epsilon}}\right)$ (cf. (113)). Using the fact that σ_i is an increasing function of θ_i (Proposition 25) and that $\Delta_\mathbf{A} \cap \mathcal{S}_0 \subset \mathcal{S}_0$, we have that $\sup_{\boldsymbol{\theta} \in \Delta_\mathbf{A} \cap \mathcal{S}_0} \sigma_i(\theta_i) \leq \sup_{\boldsymbol{\theta} \in \mathcal{S}_0} \sigma_i(\theta_i) = \sigma_i \left(\frac{1}{\lambda} \min \left(b_i, \frac{y_i - \epsilon}{\sqrt{\epsilon}} \right) \right)$ for all $i \in [m]$. This leads to the following result:

$$\alpha_{\Delta_\mathbf{A} \cap \mathcal{S}_0} = \min_{i \in [m]} -\sigma_i \left(\frac{1}{\lambda} \min \left(b_i, \frac{y_i - \epsilon}{\sqrt{\epsilon}} \right) \right) \leq \min_{i \in [m]} - \sup_{\boldsymbol{\theta} \in \Delta_\mathbf{A} \cap \mathcal{S}_0} \sigma_i(\theta_i).$$

which concludes the proof. ■

D.3.4 Strong-Concavity Bound on $\mathcal{B}(\theta, r) \cap \mathcal{S}_0$

Proposition 28. *Let $\mathcal{S}_0$ be the set defined in Proposition 26. Then, for any ball $\mathcal{B}(\boldsymbol{\theta}, r)$ with $\boldsymbol{\theta} \in \mathrm{dom}(D_\lambda) \cap \mathcal{S}_0$, the dual function D_λ as defined in (109) is $\alpha_{\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0}$ strongly concave on $\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0$ with constant:*

$$\alpha_{\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0} = \min_{i \in [m]} -\sigma_i(d_i/\lambda) \tag{120}$$

$$\text{with} \ \ d_i = \min \left(\lambda(\theta_i + r), \ b_i, \ \frac{y_i - \epsilon}{\sqrt{\epsilon}} \right). \tag{121}$$

where the b_i are quantities that define $\mathcal{S}_0$.

Proof From Proposition 10 (with $\mathcal{I} = \emptyset$) we have to prove that

$$\alpha_{\mathcal{B}(\boldsymbol{\theta},r)\cap\mathcal{S}_0} \leq \min_{i\in[m]} - \sup_{\boldsymbol{\theta}'\in\mathcal{B}(\boldsymbol{\theta},r)\cap\mathcal{S}_0} \sigma_i(\theta_i'). \tag{122}$$

By definition of $\mathcal{S}_0$ in (113) we have that

$$\boldsymbol{\theta}' \in \mathcal{B}(\boldsymbol{\theta},r)\cap\mathcal{S}_0 \iff \forall i\in[m], \theta_i' \in \left[\theta_i - r,\ \min\left(\theta_i + r,\ \frac{b_i}{\lambda},\ \frac{y_i-\epsilon}{\lambda\sqrt{\epsilon}}\right)\right]. \tag{123}$$

Combining that with the fact that σ_i is an increasing function of θ_i (Proposition 25), we obtain

$$\sup_{\boldsymbol{\theta}'\in\mathcal{B}(\boldsymbol{\theta},r)\cap\mathcal{S}_0} \sigma_i(\theta_i') = \sigma_i\left(\min\left(\theta_i + r,\ \frac{b_i}{\lambda},\ \frac{y_i-\epsilon}{\lambda\sqrt{\epsilon}}\right)\right), \tag{124}$$

which completes the proof. ■

D.3.5 Dual Update

The so-called generalized residual (Ndiaye et al., 2017) is given by

$$\boldsymbol{\rho}(\mathbf{x}) := -\nabla F(\mathbf{A}\mathbf{x}) = \frac{\mathbf{y}}{\sqrt{\mathbf{A}\mathbf{x}+\epsilon}} - \sqrt{\mathbf{A}\mathbf{x}+\epsilon}$$

and the dual feasible point $\boldsymbol{\theta} \in \Delta_{\mathbf{A}}$ defined via scaling in equation (24) becomes:

$$\boldsymbol{\theta} = \boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda) = \boldsymbol{\Xi}\left(\frac{1}{\lambda}\left(\frac{\mathbf{y}}{\sqrt{\mathbf{A}\mathbf{x}+\epsilon}} - \sqrt{\mathbf{A}\mathbf{x}+\epsilon}\right)\right) \tag{125}$$

When using Algorithms 2 and 3, the computation of the dual feasible point should also take into account $\mathcal{S}_0$. A dual feasible point $\boldsymbol{\Theta}(\mathbf{x}) \in \Delta_{\mathbf{A}} \cap \mathcal{S}_0$ can be computed as shown in Proposition 29.

Proposition 29. *Let $\mathcal{S}_0$ be the set defined in Proposition 26 and $\mathbf{A} \in \mathbb{R}_+^{m\times n}$. Let $\mathbf{x} \in \mathbb{R}_+^n$ be a primal feasible point and $\boldsymbol{\rho}(\mathbf{x}) \in \mathbb{R}^m$ the corresponding residual. Then $\boldsymbol{\Theta}(\mathbf{x}) \in \mathbb{R}^m$ defined as follows*

$$[\boldsymbol{\Theta}(\mathbf{x})]_i = \min\left([\boldsymbol{\Xi}\left(\boldsymbol{\rho}(\mathbf{x})/\lambda\right)]_i,\ \frac{b_i}{\lambda},\ \frac{y_i-\epsilon}{\lambda\sqrt{\epsilon}}\right) \tag{126}$$

is such that $\boldsymbol{\Theta}(\mathbf{x}) \in \Delta_{\mathbf{A}} \cap \mathcal{S}_0$. Moreover, we have $\boldsymbol{\Theta}(\mathbf{x}) \to \boldsymbol{\theta}^\star$ as $\mathbf{x} \to \mathbf{x}^\star$

Proof First, note that $\boldsymbol{\Theta}(\mathbf{x}) \in \mathcal{S}_0$ by construction, since $\lambda[\boldsymbol{\Theta}(\mathbf{x})]_i \leq b_i$ and $\lambda[\boldsymbol{\Theta}(\mathbf{x})]_i \leq \frac{y_i-\epsilon}{\sqrt{\epsilon}}$ for all $i \in [m]$. To show that $\boldsymbol{\Theta}(\mathbf{x}) \in \Delta_{\mathbf{A}} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \mathbf{A}^\mathsf{T}\boldsymbol{\theta} \leq \mathbf{1}\}$, we use the fact that $\boldsymbol{\Theta}(\mathbf{x}) \leq \boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda)$ and, multiplying both sides by $\mathbf{A}^\mathsf{T}$, we obtain $\mathbf{A}^\mathsf{T}\boldsymbol{\Theta}(\mathbf{x}) \leq \mathbf{A}^\mathsf{T}\boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda) \leq \mathbf{1}$ using the definition of $\boldsymbol{\Xi}$ (see (25)) for the last inequality.

Moreover, because $F \in C^1$, we have from Lemma 9 that $\boldsymbol{\Xi}(-\boldsymbol{\rho}(\mathbf{x})/\lambda) \to \boldsymbol{\theta}^\star$ as $\mathbf{x} \to \mathbf{x}^\star$. Then, because $\boldsymbol{\theta}^\star \in \mathcal{S}_0$, we have that $\min\left([\boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x}^\star)/\lambda)]_i,\ \frac{b_i}{\lambda},\ \frac{y_i-\epsilon}{\lambda\sqrt{\epsilon}}\right) = [\boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x}^\star)/\lambda)]_i = \theta_i^\star$ for all $i \in [m]$. Hence, by continuity of the operator $\boldsymbol{\Theta}$ in (126) (which is the min of two continuous functions), we conclude that $\boldsymbol{\Theta}(\mathbf{x}) \to \boldsymbol{\theta}^\star$ as $\mathbf{x} \to \mathbf{x}^\star$. ■

Remark 30. *Because $\boldsymbol{\theta} = \boldsymbol{\Theta}(\mathbf{x})$ in eq.* (126) *is no longer a simple scaling of $\boldsymbol{\rho}(\mathbf{x})$, the quantity $\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}$ required by the screening test cannot be obtained directly from $\mathbf{a}_j^\mathsf{T}\boldsymbol{\rho}(\mathbf{x})$ (which is often available from the solver's update step). Obviously, calculating $\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}$ from scratch is always an option, which may remain interesting if the computational savings provided by screening compensates this computational overhead on the screening test computation. Another option (the one adopted in our experiments) is to use $\boldsymbol{\theta}' = \boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda)$ in eq.* (125) *for the purpose of the screening test only. This way, $\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}'$ is just a scaling of $\mathbf{a}_j^\mathsf{T}\boldsymbol{\rho}(\mathbf{x})$ and the screening test remains safe since $\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}' \geq \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}$ (indeed, $\theta'_i = [\boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda)]_i \geq [\boldsymbol{\Theta}(\mathbf{x})]_i = \theta_i \ \forall i \in [m]$).*

D.4 Kullback-Leibler Divergence (β-Divergence with $\beta = 1$)

The data fidelity term is given by the Kullback Leibler divergence between the input signal $\mathbf{y} \in \mathbb{R}_+^m$ and its reconstruction $\mathbf{Ax}$, i.e.

$$F(\mathbf{Ax}) = \mathbf{y}^\mathsf{T} \log\left(\frac{\mathbf{y}}{\mathbf{Ax}+\epsilon}\right) + \mathbf{1}^\mathsf{T}(\mathbf{Ax}+\epsilon-\mathbf{y}) \tag{127}$$

$$f_i([\mathbf{Ax}]_i) = y_i \log\left(\frac{y_i}{[\mathbf{Ax}]_i+\epsilon}\right) + [\mathbf{Ax}]_i + \epsilon - y_i \tag{128}$$

where, just like in the previous case of $\beta \in (1,2)$, we introduce an ϵ-smoothing factor ($\epsilon > 0$) on the second variable. Also, similarly to Appendix D.3, we set $\mathcal{C} = \mathbb{R}_+^n$ and we consider that $\mathbf{A} \in \mathbb{R}_+^{m\times n}$.

Its first derivative is given by:

$$\nabla F(\mathbf{z}) = -\frac{\mathbf{y}}{\mathbf{z}+\epsilon} + \mathbf{1} \qquad\qquad f_i'(z_i) = -\frac{y_i}{z_i+\epsilon} + 1 \tag{129}$$

We then deduce from Theorem 1 that the first-order optimality condition (7) (primal-dual link) is given by:

$$\lambda\boldsymbol{\theta}^\star = \frac{\mathbf{y}}{\mathbf{Ax}^\star+\epsilon} - \mathbf{1} \qquad\qquad \lambda\theta_i^\star = \frac{y_i}{[\mathbf{Ax}^\star]_i+\epsilon} - 1, \quad \forall i \in [m] \tag{130}$$

The maximum regularization parameter $\lambda_{\max}$ is obtained by subtituting $\nabla F(\mathbf{0}) = \frac{\epsilon-\mathbf{y}}{\epsilon}$ in (20):

$$\lambda_{\max} = \max(\mathbf{A}^\mathsf{T}(\mathbf{y}-\epsilon)/\epsilon) \tag{131}$$

Proposition 31. *The Fenchel conjugate of the KL-divergence in equation* (127) *is given by $F^*(\mathbf{u}) = \sum_{i=1}^m f_i^*(u_i)$ where*

$$f_i^*(u) = -y_i \log(1-u) - \epsilon u \tag{132}$$

with $\mathrm{dom}(f_i^*) = (-\infty, 1)$.

Proof The Fenchel conjugate f_i^* of the scalar function $f_i = d_{\mathrm{KL}}$ is given by

$$\begin{aligned} f_i^*(u) &= \sup_{z\in\mathbb{R}} \; zu - f_i(z) \\ &= \sup_{z+\epsilon\geq 0} \; zu - y_i \log\left(\frac{y_i}{z+\epsilon}\right) + y_i - z - \epsilon \\ &= y_i - \epsilon + \sup_{z+\epsilon\geq 0} \underbrace{z(u-1) - y_i \log\left(\frac{y_i}{z+\epsilon}\right)}_{\varphi(z)} \end{aligned} \tag{133}$$

To solve this supremum problem, we distinguish two cases.

- When $y_i > 0$, the result follows by solving $\varphi'(z^\star) = u - 1 + \frac{y_i}{z^\star+\epsilon} = 0 \iff z^\star = \frac{y_i}{1-u} - \epsilon$, which is a global maximum as φ is strictly concave with $\varphi''(z) = -\frac{y_i}{(z+\epsilon)^2} < 0$. Plugging $z^\star$ into (133) we obtain:

$$f_i^*(u) = -y_i \log(1-u) - \epsilon u. \tag{134}$$

  with domain given by $\mathrm{dom}(f_i^*) = (-\infty, 1)$.

- When $y_i = 0$, one can easily verify that

$$f_i^*(u) = -\epsilon + \sup_{z+\epsilon\geq 0} z(u-1) = \begin{cases} -\epsilon u & \text{if } u < 1 \\ +\infty & \text{otherwise.} \end{cases} \tag{135}$$

  with the supremum being attained at $z^\star = -\epsilon$ in the case $u < 1$. Then, note that the same result is retrieved by taking $y_i = 0$ in equation (134), which can therefore be used in all cases.

This completes the proof. ■

The dual function $D_\lambda(\boldsymbol{\theta}) = -\sum_{i=1}^m f_i^*(-\lambda\theta_i)$ is given by:

$$D_\lambda(\boldsymbol{\theta}) = \sum_{i=1}^m y_i \log(1+\lambda\theta_i) - \epsilon\lambda\theta_i \tag{136}$$

with domain given by $\mathrm{dom}(D_\lambda) = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\theta} \geq -\mathbf{1}/\lambda\}$. Then, we get from Theorem 1, $\mathcal{C} = \mathbb{R}_+^n$, and (19) (dual norm of the ℓ_1-norm) that

$$\Delta_{\mathbf{A}} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \mathbf{A}^\mathsf{T}\boldsymbol{\theta} \leq \mathbf{1}, \boldsymbol{\theta} \geq -\mathbf{1}/\lambda\} \tag{137}$$

The Hessian $\nabla^2 D_\lambda(\boldsymbol{\theta})$ and corresponding eigenvalues $\sigma_i(\theta_i)$ are given by

$$\nabla^2 D_\lambda(\boldsymbol{\theta}) = \mathrm{Diag}\left([\sigma_i(\theta_i)]_{i\in[m]}\right), \qquad \sigma_i(\theta_i) = -\frac{\lambda^2 y_i}{(1+\lambda\theta_i)^2} \tag{138}$$

Although the eigenvalues are all non-positive, they tend to zero as $|\theta_i|$ tends to infinity. Moreover, $\sigma_i(\theta_i)$ also vanishes when $y_i = 0$. Therefore $D_\lambda(\boldsymbol{\theta})$ is not globally strongly concave and the standard Gap Safe approach cannot be applied in this case.

In the following sections we find local strong concavity bounds that allows us to deploy the proposed screening strategy (Algorithms 2 and 3).

Proposition 32. *The eigenvalues $\sigma_i(\theta_i)$ in* (138) *are an increasing (resp. strictly increasing) function of θ_i for any $y_i \geq 0$ and $\theta_i \geq -1/\lambda$ (resp. $y_i > 0$ and $\theta_i > -1/\lambda$) .*

Proof Indeed, its first derivative is non-negative

$$\sigma_i'(\theta_i) = \lambda^3 \frac{2y_i}{(1+\lambda\theta_i)^3} \geq 0 \tag{139}$$

since $\mathbf{y} \in \mathbb{R}^m_+$ and $\theta_i \geq -1/\lambda \implies 1 + \lambda\theta_i \geq 0$. ■

D.4.1 Validity of the Set $\mathcal{S}_0$

In order to be able to derive local strong concavity bounds (see Sections D.4.2 and D.4.3 hereafter), we need to use the set $\mathcal{S}_0$ below (see discussion in Section 4.2.2). To comply with Theorem 5, we need to show that $\boldsymbol{\theta}^\star \in \mathcal{S}_0$.

Proposition 33. *Let $\mathcal{I}_0 = \{i \in [m] \mid y_i = 0\}$. Then, for any $\mathbf{A} \in \mathbb{R}^{m\times n}_+$, $\mathbf{y} \in \mathbb{R}_+$, the set*

$$\mathcal{S}_0 = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\theta}_{\mathcal{I}_0} = -\mathbf{1}/\lambda\} \tag{140}$$

is such that $\boldsymbol{\theta}^\star \in \mathcal{S}_0$.

Proof Optimality condition (130) directly implies that the dual solution takes the value $\theta_i^\star = -1/\lambda$ for all coordinates $i \in \mathcal{I}_0$, i.e. all coordinates i such that $y_i = 0$. ■

With that in hand, we are now able to compute local strong concavity bounds on both $\Delta_{\mathbf{A}} \cap \mathcal{S}_0$ and $\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0$.

D.4.2 Strong-Concavity Bound on $\Delta_A \cap \mathcal{S}_0$

Proposition 34. *Let $\mathcal{S}_0$ be the set defined in Proposition 33 and $\mathbf{A} \in \mathbb{R}^{m\times n}_+$ with no all-zero row. Then, the dual function D_λ as defined in* (136) *is $\alpha_{\Delta_{\mathbf{A}}\cap\mathcal{S}_0}$ strongly concave on $\Delta_{\mathbf{A}} \cap \mathcal{S}_0$ with:*

$$\alpha_{\Delta_{\mathbf{A}}\cap\mathcal{S}_0} = \lambda^2 \min_{i\in\mathcal{I}_0^\complement} \frac{y_i}{\left(\min_{\{j\in[n] \mid a_{ij}\neq 0\}} \left(\frac{\lambda + \|\mathbf{a}_j\|_1}{a_{ij}}\right)\right)^2} \tag{141}$$

Proof From Proposition 10 (with $\mathcal{I} = \mathcal{I}_0$) we have to prove that

$$\alpha_{\Delta_{\mathbf{A}}\cap\mathcal{S}_0} \leq \min_{i\in\mathcal{I}_0^\complement} - \sup_{\boldsymbol{\theta}\in\Delta_{\mathbf{A}}\cap\mathcal{S}_0} \sigma_i(\theta_i) \tag{142}$$

where $\sigma_i(\theta_i) = -\frac{\lambda^2 y_i}{(1+\lambda\theta_i)^2}$ denotes the i-th eigenvalue of $\nabla^2 D_\lambda(\boldsymbol{\theta})$. Let $\boldsymbol{\theta} \in \Delta_{\mathbf{A}} \cap \mathcal{S}_0 = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \mathbf{A}^\mathsf{T}\boldsymbol{\theta} \leq \mathbf{1},\ \boldsymbol{\theta} \geq -1/\lambda,\ \boldsymbol{\theta}_{\mathcal{I}_0} = -\mathbf{1}/\lambda\}$. Then, for all $i \in [m]$, we have

$$1 \geq \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta} = \sum_{i'=1}^m a_{i'j}\theta_{i'} = a_{ij}\theta_i + \sum_{i'\neq i} a_{i'j}\theta_{i'} \geq a_{ij}\theta_i - \sum_{i'\neq i} a_{i'j}\frac{1}{\lambda},\ \forall j \in [n], \tag{143}$$

using the facts that $\mathbf{A} \in \mathbb{R}_+^{m\times n}$ and $\theta_i \geq -1/\lambda$, for all $i \in [m]$. Reorganising the terms in the above inequalities and using the definition of $\mathrm{dom}(D_\lambda)$ (with $\theta_i \geq -1/\lambda$) we obtain, for all $i \in [m]$,

$$-\frac{1}{\lambda} \leq \theta_i \leq \frac{1+\sum_{i'\neq i} a_{i'j}\frac{1}{\lambda}}{a_{ij}} = \frac{\lambda + \|\mathbf{a}_j\|_1}{\lambda a_{ij}} - \frac{1}{\lambda}, \quad \forall j \in [n] \text{ s.t. } a_{ij} \neq 0 \tag{144}$$

$$\iff \quad 0 \leq 1 + \lambda\theta_i \leq \min_{\{j\in[n] \mid a_{ij}\neq 0\}} \left(\frac{\lambda + \|\mathbf{a}_j\|_1}{a_{ij}}\right) \tag{145}$$

where our assumption that $\mathbf{A}$ has no all-zero row implies that the set $\{j \in [n] \mid a_{ij} \neq 0\}$ is non-empty. As σ_i is an increasing function of θ_i (Proposition 32), we get that the sup in (142) is attained for $1 + \lambda\theta_i = \min_{\{j\in[n] \mid a_{ij}\neq 0\}} \left(\frac{\lambda+\|\mathbf{a}_j\|_1}{a_{ij}}\right)$, which completes the proof. ■

D.4.3 Strong-Concavity Bound on $\mathcal{B}(\theta, r) \cap \mathcal{S}_0$

Proposition 35. *Let $\mathcal{S}_0$ be the set defined in Proposition 33. Then, for any ball $\mathcal{B}(\boldsymbol{\theta}, r)$ with $\boldsymbol{\theta} \in \mathrm{dom}(D_\lambda) \cap \mathcal{S}_0$, the dual function D_λ as defined in (136) is $\alpha_{\mathcal{B}(\boldsymbol{\theta},r)\cap\mathcal{S}_0}$ strongly concave on $\mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0$ with:*

$$\alpha_{\mathcal{B}(\boldsymbol{\theta},r)\cap\mathcal{S}_0} = \lambda^2 \min_{i\in\mathcal{I}_0^\complement} \frac{y_i}{(1+\lambda(\theta_i + r))^2}. \tag{146}$$

Proof From Proposition 10 (with $\mathcal{I} = \mathcal{I}_0$) we have to prove that

$$\alpha_{\mathcal{B}(\boldsymbol{\theta},r)\cap\mathcal{S}_0} \leq \min_{i\in\mathcal{I}_0^\complement} - \sup_{\boldsymbol{\theta}'\in\mathcal{B}(\boldsymbol{\theta},r)\cap\mathcal{S}_0} \sigma_i(\theta_i') \tag{147}$$

$$= \min_{i\in\mathcal{I}_0^\complement} - \sup_{\boldsymbol{\theta}'\in\mathcal{B}(\boldsymbol{\theta},r)\cap\mathcal{S}_0} -\frac{\lambda^2 y_i}{(1+\lambda\theta_i')^2} \tag{148}$$

$$= \lambda^2 \min_{i\in\mathcal{I}_0^\complement} \inf_{|\theta_i'-\theta_i|\leq r} \frac{y_i}{(1+\lambda\theta_i')^2} \tag{149}$$

$$= \lambda^2 \min_{i\in\mathcal{I}_0^\complement} \frac{y_i}{(1+\lambda(\theta_i + r))^2} \tag{150}$$

where we used the facts that $y_i > 0$ and $1 + \lambda\theta_i \geq 0$ (using the definition of $\mathrm{dom}(D_\lambda)$ since $\boldsymbol{\theta} \in \mathrm{dom}(D_\lambda)$) to conclude that the infimum is attained for $\theta_i + r$ rather than $\theta_i - r$. ■

D.4.4 Dual Update

The generalized residual w.r.t. a primal estimate $\mathbf{x}$ is given by

$$\boldsymbol{\rho}(\mathbf{x}) := -\nabla F(\mathbf{A}\mathbf{x}) = \frac{\mathbf{y}}{\mathbf{A}\mathbf{x} + \epsilon} - \mathbf{1}$$

and the dual feasible point $\boldsymbol{\theta} \in \Delta_{\mathbf{A}}$ obtained via scaling in equation (24) is given by:

$$\boldsymbol{\theta} = \boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda) = \boldsymbol{\Xi}\left(\frac{1}{\lambda}\left(\frac{\mathbf{y}}{\mathbf{A}\mathbf{x}+\epsilon} - \mathbf{1}\right)\right) \tag{151}$$

However, this is not sufficient in order to apply Algorithms 2 and 3. Indeed, one needs to compute a dual point $\boldsymbol{\theta} \in \Delta_{\mathbf{A}} \cap \mathcal{S}_0$.

Proposition 36. *Let $\mathcal{S}_0$ be the set defined in Proposition 33 and $\mathbf{A} \in \mathbb{R}_+^{m \times n}$. Let $\mathbf{x} \in \mathbb{R}_+^n$ be a primal feasible point and $\boldsymbol{\rho}(\mathbf{x}) \in \mathbb{R}^m$ the corresponding residual. Then $\boldsymbol{\Theta}(\mathbf{x}) \in \mathbb{R}^m$ defined as follows*

$$[\boldsymbol{\Theta}(\mathbf{x})]_i = \begin{cases} [\boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda)]_i & \text{if } \ i \in \mathcal{I}_0^\complement \\ -\frac{1}{\lambda} & \text{if } \ i \in \mathcal{I}_0 \end{cases} \tag{152}$$

is such that $\boldsymbol{\Theta}(\mathbf{x}) \in \Delta_{\mathbf{A}} \cap \mathcal{S}_0$. Moreover, we have $\boldsymbol{\Theta}(\mathbf{x}) \to \boldsymbol{\theta}^\star$ as $\mathbf{x} \to \mathbf{x}^\star$.

Proof By construction, we have $\boldsymbol{\Theta}(\mathbf{x}) \in \mathcal{S}_0$. It remains to show that $\boldsymbol{\Theta}(\mathbf{x}) \in \Delta_{\mathbf{A}} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \mathbf{A}^\mathsf{T}\boldsymbol{\theta} \leq \mathbf{1}, \lambda\boldsymbol{\theta} \geq -\mathbf{1}\}$, where the second inequality corresponds to the domain of the dual function. Because $\mathbf{A} \in \mathbb{R}_+^{m \times n}$, $\mathbf{x} \in \mathbb{R}_+^n$, and $\mathbf{y} \in \mathbb{R}_+^m$, we have that $\boldsymbol{\rho}(\mathbf{x})/\lambda \geq -\mathbf{1}/\lambda$. From the definition of the scaling $\boldsymbol{\Xi}$ in (25), –the scaling factor always being on the interval $(0, 1]$– we obtain that $\boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda) \geq \boldsymbol{\Theta}(\mathbf{x}) \geq -\mathbf{1}/\lambda$. Multiplying the left and right hand side of the first inequality by $\mathbf{A}^\mathsf{T}$, we obtain $\mathbf{A}^\mathsf{T}\boldsymbol{\Theta}(\mathbf{x}) \leq \mathbf{A}^\mathsf{T}\boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda) \leq \mathbf{1}$ (by the definition of $\boldsymbol{\Xi}$ in (25) for the second inequality) which shows that $\boldsymbol{\Theta}(\mathbf{x}) \in \Delta_{\mathbf{A}}$. Moreover, from the optimality condition (130) we have $[\boldsymbol{\Theta}(\mathbf{x})]_{\mathcal{I}_0} = \boldsymbol{\theta}^\star_{\mathcal{I}_0}$ regardless of $\mathbf{x}$ and for the remaining coordinates we obtain from Lemma 9 (since $F \in C^1$) that $\boldsymbol{\Xi}(-\boldsymbol{\rho}(\mathbf{x})/\lambda) \to \boldsymbol{\theta}^\star$ as $\mathbf{x} \to \mathbf{x}^\star$ which concludes the proof that $\boldsymbol{\Theta}(\mathbf{x}) \to \boldsymbol{\theta}^\star$ as $\mathbf{x} \to \mathbf{x}^\star$. ■

D.4.5 Improved Screening Test

An improved screening test can be defined on a Gap Safe sphere when intersected with $\mathcal{S}_0$.

Proposition 37. *Let $\mathcal{I}_0 = \{i \in [m] : y_i = 0\}$. Let $\boldsymbol{\theta} \in \mathcal{S}_0$ and $r > 0$ be such that $\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r)$. Then,*

$$\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta} + r\|[\mathbf{a}_j]_{\mathcal{I}_0}\|_2 < 1 \quad \implies \quad x_j^\star = 0. \tag{153}$$

Proof First of all, one can see from (130) that the dual solution takes the value $\theta_i^\star = -1/\lambda$ for all coordinates $i \in \mathcal{I}_0$. Hence, by definition of $\mathcal{S}_0$, we have $\boldsymbol{\theta}^\star \in \mathcal{S}_0$ and thus $\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0$. Therefore, from Proposition 3, we have that

$$\sup_{\boldsymbol{\xi} \in \mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0} \mathbf{a}_j^\mathsf{T}\boldsymbol{\xi} < 1 \implies \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}^\star < 1 \implies x_j^\star = 0.$$

Finally, we have

$$\begin{aligned} \sup_{\boldsymbol{\xi} \in \mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}_0} \mathbf{a}_j^\mathsf{T}\boldsymbol{\xi} &= \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta} + r \sup_{\mathbf{u} \in \mathcal{B}(\mathbf{0}, 1), \mathbf{u}_{\mathcal{I}_0} = \mathbf{0}} \mathbf{a}_j^\mathsf{T}\mathbf{u} \\ &= \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta} + r\|[\mathbf{a}_j]_{\mathcal{I}_0}\|_2 \end{aligned}$$

which completes the proof. ■

Remark 38. *When using $\boldsymbol{\theta} = \boldsymbol{\Theta}(\mathbf{x})$ in eq. (152), the quantity $\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}$ required for the screening test can be obtained with mild computational effort from $\mathbf{a}_j^\mathsf{T}\boldsymbol{\rho}(\mathbf{x})$ (which is usually calculated in the solver's update step) even if $\boldsymbol{\Theta}(\mathbf{x})$ is no longer a simple scaled version of $\boldsymbol{\rho}(\mathbf{x})$. Indeed, we have:*

$$\begin{aligned}\mathbf{a}_j^\mathsf{T}\boldsymbol{\Theta}(\mathbf{x}) &= \sum_{i' \in \mathcal{I}_0^\complement} a_{i'j}[\boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda)]_{i'} + \sum_{i\in\mathcal{I}_0} a_{ij}\left(-\frac{1}{\lambda}\right) \\ &= \mathbf{a}_j^\mathsf{T}\boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda) - \sum_{i\in\mathcal{I}_0} a_{ij}\left(\frac{1}{\lambda} + [\boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda)]_i\right) \\ &= s\ \mathbf{a}_j^\mathsf{T}\boldsymbol{\rho}(\mathbf{x})/\lambda - \frac{1}{\lambda}\,(1-s)\,\|[\mathbf{a}_j]_{\mathcal{I}_0}\|_1\end{aligned}$$

where we denoted s the scaling factor in $\boldsymbol{\Xi}$, such that $\boldsymbol{\Xi}(\mathbf{z}) = s\mathbf{z}$, and in the last equality we used the fact that $[\boldsymbol{\Xi}(\boldsymbol{\rho}(\mathbf{x})/\lambda)]_{\mathcal{I}_0} = -s\frac{\mathbf{1}}{\lambda}$. The final expression can be computed efficiently, since the norms $\|[\mathbf{a}_j]_{\mathcal{I}_0}\|_1$ can be precomputed and all remaining operations are scalar sums and multiplications.

References

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1): 1–106, 2012. ISSN 1935-8237. doi: 10.1561/2200000015. URL `http://dx.doi.org/10.1561/2200000015`.

Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998. ISSN 00063444. URL `http://www.jstor.org/stable/2337385`.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 1st edition, 2011. ISBN 1441994661.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, Jan 2009. doi: 10.1137/080716542. URL `http://dx.doi.org/10.1137/080716542`.

Antoine Bonnefoy, Valentin Emiya, Liva Ralaivola, and Rémi Gribonval. Dynamic screening: Accelerating first-order algorithms for the lasso and group-lasso. *IEEE Transactions on Signal Processing*, 63(19):5121–5132, Oct 2015. ISSN 1053-587X. doi: 10.1109/TSP.2015.2447503.

J.M. Borwein and A.S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS books in mathematics. Springer, 2000. ISBN 9780387989402. URL `https://www.springer.com/gp/book/9780387295701`.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. ISBN 978-3-642-20191-2. doi: 10.1007/978-3-642-20192-9. URL `http://dx.doi.org/10.1007/978-3-642-20192-9`. Methods, theory and applications.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, May 2011. ISSN 1573-7683. doi: 10.1007/s10851-010-0251-1. URL https://doi.org/10.1007/s10851-010-0251-1.

Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994. ISSN 00401706. URL http://www.jstor.org/stable/1269949.

C. F. Dantas and R. Gribonval. Stable safe screening and structured dictionaries for faster ℓ_1 regularization. *IEEE Transactions on Signal Processing*, 67(14):3756–3769, July 2019. ISSN 1053-587X. doi: 10.1109/TSP.2019.2919404.

Cassio F. Dantas, Emmanuel Soubies, and Cédric Févotte. Safe screening for sparse regression with the kullback-leibler divergence. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, June 2021.

Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. *Pacific Journal of Optimization*, 8(4):667–698, Oct 2012. Special Issue on Conic Optimization.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911, Nov 2008. doi: 10.1111/j.1467-9868.2008.00674.x. URL https://doi.org/10.1111/j.1467-9868.2008.00674.x.

O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: safer rules for the lasso. In *International Conference on Machine Learning*, volume 37, pages 333–342, July 2015.

Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, Sep 2011. ISSN 0899-7667. doi: 10.1162/NECO_a_00168. URL https://doi.org/10.1162/NECO_a_00168.

Cédric Févotte, Emmanuel Vincent, and Alexey Ozerov. Single-channel audio source separation with NMF: divergences, constraints and algorithms. In *Audio Source Separation*. Springer, March 2018. URL https://hal.inria.fr/hal-01631185.

Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i01. URL https://www.jstatsoft.org/v033/i01.

C. Févotte and N. Dobigeon. Nonlinear hyperspectral unmixing with robust nonnegative matrix factorization. *IEEE Transactions on Image Processing*, 24(12):4810–4819, 2015. doi: 10.1109/TIP.2015.2468177.

A. Globerson, G. Chechik, F. Pereira, and N. Tishby. Euclidean Embedding of Co-occurrence Data. *The Journal of Machine Learning Research*, 8:2265–2295, 2007.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286: 531–537, 1999.

Z. T. Harmany, R. F. Marcia, and R. M. Willett. This is SPIRAL-TAP: Sparse Poisson Intensity Reconstruction ALgorithms—Theory and Practice. *IEEE Transactions on Image Processing*, 21(3):1084–1096, March 2012. ISSN 1941-0042. doi: 10.1109/TIP.2011.2168410.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer Berlin Heidelberg, 1993a. doi: 10.1007/978-3-662-02796-7. URL `https://doi.org/10.1007/978-3-662-02796-7`.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms II*. Springer Berlin Heidelberg, 1993b. doi: 10.1007/978-3-662-06409-2. URL `https://doi.org/10.1007/978-3-662-06409-2`.

Cho-Jui Hsieh and Inderjit S. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, page 1064–1072, 2011. ISBN 9781450308137. doi: 10.1145/2020408.2020577. URL `https://doi.org/10.1145/2020408.2020577`.

S. Jia and Y. Qian. Spectral and spatial complexity-based hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 45(12):3867–3879, 2007. doi: 10.1109/TGRS.2007.898443.

Tyler Johnson and Carlos Guestrin. Blitz: A principled meta-algorithm for scaling sparse optimization. In *International Conference on Machine Learning*, pages 1171–1179, 2015.

Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, pages 556–562. MIT Press, 2001. URL `http://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf`.

Jun Liu, Zheng Zhao, Jie Wang, and Jieping Ye. Safe screening with variational inequalities and its application to lasso. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32(2) of *ICML'14*, pages 289–297. JMLR.org, June 2014. URL `http://dl.acm.org/citation.cfm?id=3044805.3044925`.

A. Malti and C. Herzet. Safe screening tests for LASSO based on firmly non-expansiveness. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar 2016. doi: 10.1109/icassp.2016.7472575. URL `https://doi.org/10.1109/icassp.2016.7472575`.

M. Massias, A. Gramfort, and J. Salmon. From safe screening rules to working sets for faster Lasso-type solvers. In *NIPS Workshop on Optimization for Machine Learning*, Long Beach, USA, Dec 2017.

Mathurin Massias, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Dual extrapolation for sparse glms. *Journal of Machine Learning Research*, 21(234):1–33, 2020. URL http://jmlr.org/papers/v21/19-587.html.

Eugene Ndiaye. *Safe optimization algorithms for variable selection and hyperparameter tuning*. Thesis, Université Paris-Saclay, Oct 2018. URL https://pastel.archives-ouvertes.fr/tel-01962450.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparse-group lasso. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 388–396. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/555d6702c950ecb729a966504af0a635-Paper.pdf.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, 18(128):1–33, Nov 2017.

Kohei Ogawa, Yoshiki Suzuki, and Ichiro Takeuchi. Safe screening of non-support vectors in pathwise svm computation. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1382–1390, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/ogawa13b.html.

S. Ren, S. Huang, J. Ye, and X. Qian. Safe feature screening for generalized lasso. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2992–3006, 2018. doi: 10.1109/TPAMI.2017.2776267.

R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74 (2):245–266, Nov 2011. doi: 10.1111/j.1467-9868.2011.01004.x. URL https://doi.org/10.1111/j.1467-9868.2011.01004.x.

T. Virtanen, J. F. Gemmeke, and B. Raj. Active-set newton algorithm for overcomplete non-negative representations of audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2277–2289, 2013. doi: 10.1109/TASL.2013.2263144.

J. Wang, W. Fan, and J. Ye. Fused lasso screening rules via the monotonicity of subdifferentials. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1806–1820, 2015. doi: 10.1109/TPAMI.2014.2388203.

Jie Wang, Peter Wonka, and Jieping Ye. Scaling svm and least absolute deviations via exact data reduction. In *International Conference on Machine Learning (ICML)*, ICML'14, page II–523–II–531. JMLR.org, 2014.

Jie Wang, Peter Wonka, and Jieping Ye. Lasso screening rules via dual polytope projection. *Journal of Machine Learning Research*, 16(1):1063–1101, May 2015a. ISSN 1532-4435. URL http://jmlr.org/papers/v16/wang15a.html.

Jie Wang, Zhanqiu Zhang, and Jieping Ye. Two-layer feature reduction for sparse-group lasso via decomposition of convex sets. *Journal of Machine Learning Research*, 20(163): 1–42, 2019. URL http://jmlr.org/papers/v20/16-383.html.

Suyu Wang, Zongxiang Zhang, and Ying Wu. Spatial and spectral coordinate super resolution of hyperspectral imagery based on redundant dictionary. In *Seventh International Conference on Graphic and Image Processing*, pages 98170E–98170E. International Society for Optics and Photonics, 2015b.

Z. J. Xiang and P. J. Ramadge. Fast lasso screening tests based on correlations. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2137–2140, March 2012. doi: 10.1109/ICASSP.2012.6288334.

Z. J. Xiang, Y. Wang, and P. J. Ramadge. Screening tests for lasso problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):1008–1027, May 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2568185.

Zhen James Xiang, Hao Xu, and Peter J Ramadge. Learning sparse representations of high dimensional data on large scale dictionaries. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24, pages 900–908, Granada, Spain, Dec 2011.

F. Yanez and F. Bach. Primal-dual algorithms for non-negative matrix factorization with the Kullback-Leibler divergence. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2257–2261, 2017.

Tomoki Yoshida, Ichiro Takeuchi, and Masayuki Karasuyama. Safe triplet screening for distance metric learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2653–2662, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3220037. URL https://doi.org/10.1145/3219819.3220037.

Guo-Xun Yuan, Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. A comparison of optimization methods and software for large-scale l1-regularized linear classification. *Journal of Machine Learning Research*, 11(105):3183–3234, 2010. URL http://jmlr.org/papers/v11/yuan10c.html.

Qiang Zhou and Qi Zhao. Safe subspace screening for nuclear norm regularized least squares problems. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1103–1112, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/zhoua15.html.

Julian Zimmert, Christian Schroeder de Witt, Giancarlo Kerg, and Marius Kloft. Safe screening for support vector machines. In *NIPS 2015 Workshop on Optimization in Machine Learning (OPT)*, 2015.