Safe screening for sparse regression with the Kullback-Leibler divergence

Cassio F. Dantas, Emmanuel Soubies, Cédric Févotte

SAFE SCREENING FOR SPARSE REGRESSION WITH THE KULLBACK-LEIBLER DIVERGENCE

Cássio F. Dantas, Emmanuel Soubies, Cédric Févotte

IRIT, Université de Toulouse, CNRS, Toulouse, France

ABSTRACT

Safe screening rules are powerful tools to accelerate iterative solvers in sparse regression problems. They allow early identification of inactive coordinates (i.e., those not belonging to the support of the solution) which can thus be screened out in the course of iterations. In this paper, we extend the GAP Safe screening rule to the ℓ_1-regularized Kullback-Leibler divergence which does not fulfil the regularity assumptions made in previous works. The proposed approach is experimentally validated on synthetic and real count data sets.

***Index Terms*—** Safe screening, KL divergence, sparsity.

1. INTRODUCTION

The Poisson observation model has been used in a variety of applications where data correspond to a series of discrete events with inherently noisy measurements. Examples include: nuclear medical imaging (e.g., Positron emission tomography [1]), astronomy [2] and traffic analysis [3]. The linear case can be written as: $\mathbf{y} \sim \mathrm{Poisson}(\mathbf{A}\mathbf{x})$, where $\mathbf{y} \in \mathbb{R}_+^m$ is the observation vector, $\mathbf{x} \in \mathbb{R}_+^n$ is the signal of interest, and $\mathbf{A} \in \mathbb{R}_+^{m\times n}$ is the measurement matrix. Poisson intensities are naturally non-negative (thus the natural non-negativity constraints).

In this context, recovering the signal of interest $\mathbf{x}$ from the observations $\mathbf{y}$ amounts to the resolution of a linear inverse problem. The resulting problem is usually ill-posed as $m < n$ (number of observations smaller than the number of unknowns) which motivates the introduction of a regularization term based on some prior knowledge on the signal $\mathbf{x}$. In this paper, we are interested in sparsity prior which can be achieved via a ℓ_1-norm regularization. This leads to the following optimization problem:

$$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}_+^n} P_\lambda(\mathbf{x}) := \mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{A}\mathbf{x}) + \lambda\|\mathbf{x}\|_1 \quad (1)$$

where $P_\lambda(\mathbf{x})$ denotes the primal objective function, parameter $\lambda > 0$ controls the sparsity level of the solution, and the data-fidelity term is the generalized Kullback-Leibler (KL) divergence [4], which also corresponds to the Poisson negative log-likelihood up to irrelevant terms [5, 6]:

$$\mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{z}) = \sum_{i=1}^m y_i \log\left(\frac{y_i}{z_i+\epsilon}\right) - y_i + (z_i+\epsilon). \quad (2)$$

The smoothing constant $\epsilon \geq 0$ allows to avoid singularities around $z_i = 0$ and is common practice, see, e.g., [7].

Several algorithms have been proposed in the literature for tackling problem (1) [4, 7, 8, 9]. In this paper, we propose a variable elimination technique called *safe screening* that can accelerate most of the existing solvers.

Safe screening techniques were originally proposed for the Lasso problem [10], but have recently been extended to a wide variety of sparse-regularized problems [11, 12, 13]. Such techniques allow to identify, before and while solving the problem, coordinates which will not be part of the solution support. Doing so might significantly accelerate the problem's resolution. As soon as a coordinate j is identified as inactive, the corresponding column of the dictionary matrix $\mathbf{a}_j$ can be removed once and for all. In such approaches, there is no risk of false identification, thus the "safe" denomination.

In this paper, we extend an existing screening technique [12] to the regularized Kullback-Leibler problem (1). Even though the approach in [12] was proposed within a generalized linear model framework, it cannot be directly applied to our problem of interest, since the Kullback-Leibler function does not fulfill the regularity hypothesis in [12].

In Section 2, we go through the main technical ingredients on deriving safe screening rules for problem (1). Then, the proposed algorithm is described in Section 3 and experimental results are given in Section 4.

Notations $[n] = \{1, \dots, n\}$ is the set of n first integers and $\mathbf{1}$ is a vector of ones (with size inferred from context). $\boldsymbol{\theta}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ denotes the restriction of $\boldsymbol{\theta} \in \mathbb{R}^m$ to its components indexed by the elements of $\mathcal{I} \subseteq [m]$. $\mathcal{I}^{\complement}$ denotes the complement of set $\mathcal{I}$. Given two vectors $\mathbf{z} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{y}/\mathbf{z}$ is the entry-wise division. We denote $\nabla D_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{z}) = \nabla_{\mathbf{z}} D_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{z})$ the gradient w.r.t. the second variable $\mathbf{z} \in \mathbb{R}_+^m$ and $\mathcal{B}(\mathbf{c}, r)$ a closed ℓ_2-ball in $\mathbb{R}^m$ with center $\mathbf{c}$ and radius r.

This work is supported by the European Research Council (ERC FACTORY-CoG-6681839). Code is available at `https://github.com/cassiofragadantas`.

2. SAFE SCREENING FOR THE KL-ℓ_1 PROBLEM

2.1. The Dual Problem

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

$$\boldsymbol{\theta}^\star = \operatorname*{argmax}_{\boldsymbol{\theta}\in\mathcal{F}_{\mathbf{A}}} D_\lambda(\boldsymbol{\theta}) := \sum_{i=1}^{m} y_i \log(1+\lambda\theta_i) - \epsilon\lambda\theta_i, \quad (3)$$

where $\mathcal{F}_{\mathbf{A}} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \lambda\boldsymbol{\theta} \geq -\mathbf{1},\ \mathbf{A}^\mathsf{T}\boldsymbol{\theta} \leq \mathbf{1}\}$ *is the dual feasible set and "* $\leq$ *" is defined component-wisely. Moreover, first-order optimality conditions for a pair of solutions* $(\mathbf{x}^\star, \boldsymbol{\theta}^\star)$ *are given by:*

$$\lambda\boldsymbol{\theta}^\star = \frac{\mathbf{y}}{\mathbf{A}\mathbf{x}^\star + \epsilon} - \mathbf{1} \quad (4)$$

$$\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}^\star = \begin{cases} 1, & \text{if } x_j^\star > 0, \\ \varrho \leq 1, & \text{if } x_j^\star = 0. \end{cases} \quad (5)$$

Proof. The dual function $D_\lambda(\boldsymbol{\theta})$ is given by the Fenchel conjugate of the data-fidelity term $\mathcal{D}_{\mathrm{KL}}$ [12, 14]. The constraint set is given by the Fenchel conjugate of $\Omega(\mathbf{x}) = \|\mathbf{x}\|_1 + \mathbb{1}_{\mathbb{R}_+^m}(\mathbf{x})$, which is calculated in [13] for the nonnegative Lasso. For compactness, we integrate to the feasible set the domain of the dual function, given by $\mathrm{dom}(D_\lambda) = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\theta} \geq -\mathbf{1}/\lambda\}$.
Optimality conditions (4) and (5) are given by [12]:

$$\lambda\boldsymbol{\theta}^\star = -\nabla\mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{A}\mathbf{x}^\star)$$
$$\mathbf{A}^\mathsf{T}\boldsymbol{\theta}^\star \in \partial\|\mathbf{x}\|_1 + \partial\mathbb{1}_{\mathbb{R}_+^m}(\mathbf{x})$$

Calculation details for the latter equation are given in [13]. Finally, it is worth mentioning that, even for $\epsilon = 0$, (4) is well-defined as $[\mathbf{A}\mathbf{x}^\star]_i = 0 \implies y_i = 0$, in which case we use the convention that $0/0 = 0$. Indeed, when $\epsilon = 0$, if $y_i \neq 0$ and $[\mathbf{A}\mathbf{x}]_i = 0$ for any $i \in [m]$, then $\mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{A}\mathbf{x}) = +\infty$ which cannot be optimal and, hence, $y_i \neq 0 \implies [\mathbf{A}\mathbf{x}^\star]_i \neq 0$ as desired. □

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

$$\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}^\star < 1 \implies x_j^\star = 0, \quad (6)$$

for any primal solution $\mathbf{x}^\star$. In other words, the support of primal solutions $\mathbf{x}^\star$ is characterized by the knowledge of the dual solution. The promise of safe screening is then to exploit this property in order to reduce the size of the primal problem (and accelerate its resolution) by screening out the coordinates that are inactive ($x_j^\star = 0$).

2.2. Safe Sphere

Because the dual solution $\boldsymbol{\theta}^\star$ is not known in advance, the condition (6) cannot be used in practice. However, we can derive a more restrictive—yet practical—sufficient condition for eliminating coordinates of the primal problem.

Proposition 1 (KL-ℓ_1 Safe Screening Rule)**.** *Let* $\mathcal{I} = \{i \in [m] : y_i = 0\}$ *and* $\mathcal{S} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\theta}_\mathcal{I} = -\mathbf{1}/\lambda\}$. *Let* $\boldsymbol{\theta} \in \mathcal{S}$ *and* $r > 0$ *be such that* $\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r)$. *Then,*

$$\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta} + r\|[\mathbf{a}_j]_{\mathcal{I}^\complement}\|_2 < 1 \quad \implies \quad x_j^\star = 0. \quad (7)$$

Proof. First of all, one can see from (4) that the dual solution takes the value $\theta_i^\star = -1/\lambda$ for all coordinates $i \in \mathcal{I}$. Hence, by the definition of $\mathcal{S}$, we have $\boldsymbol{\theta}^\star \in \mathcal{S}$ and thus $\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}$. It follows from (5) that

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

Finally, we have

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

which completes the proof. □

Hence, if one can find a ball $\mathcal{B}(\boldsymbol{\theta}, r)$ containing $\boldsymbol{\theta}^\star$ (we say that $\mathcal{B}(\boldsymbol{\theta}, r)$ is a *safe region*), Proposition 1 allows to safely identify inactive variables of $\mathbf{x}^\star$. This is possible when the dual function is α-strongly concave [12]. Indeed, given a primal-dual feasible pair $(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}_+^n \times \mathcal{F}_{\mathbf{A}}$, the authors in [12] have shown that $\mathcal{B}(\boldsymbol{\theta}, r)$ with $r = \sqrt{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})/\alpha}$, where $\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta}) := P_\lambda(\mathbf{x}) - D_\lambda(\boldsymbol{\theta})$ denotes the duality gap, is a safe region (GAP Safe Sphere). Unfortunately, for the KL divergence, the dual objective function D_λ in (3) is not globally strongly concave. However, D_λ is locally strongly concave and we shall show that this is sufficient to derive a safe region $\mathcal{B}(\boldsymbol{\theta}, r)$.

Theorem 2 (KL-ℓ_1 GAP Safe Sphere)**.** *Let* $\mathcal{I}$ *and* $\mathcal{S}$ *be defined as in Proposition 1.Let* $(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}_+^n \times (\mathcal{F}_{\mathbf{A}} \cap \mathcal{S})$ *be a primal-dual feasible pair, and assume that* $\mathrm{rank}(\mathbf{A}) = \min(m, n)$. *Then, for*

$$r = \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\bar{\alpha}}} \quad (8)$$

$$\bar{\alpha} = \lambda^2 \min_{i\in\mathcal{I}^\complement} \frac{y_i}{(1 + \max(\|\mathbf{A}\|_1, \lambda)\|\mathbf{a}_i^\dagger\|_1)^2}, \quad (9)$$

we have $\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r)$ *(i.e.,* $\mathcal{B}(\boldsymbol{\theta}, r)$ *is a safe region). In* (9) $\mathbf{a}_i^\dagger$ *denotes the* i*-th column of* $\mathbf{A}^\dagger \in \mathbb{R}^{n\times m}$*, the right pseudo-inverse of* $\mathbf{A}$ *such that* $\mathbf{A}\mathbf{A}^\dagger = \mathbf{I}_m$*, and* $\|\mathbf{A}\|_1$ *is the maximum absolute column sum of the matrix* $\mathbf{A}$.

Proof. Because by definition and assumption we have both $\boldsymbol{\theta}^\star \in \mathcal{F}_{\mathbf{A}} \cap \mathcal{S}$ and $\boldsymbol{\theta} \in \mathcal{F}_{\mathbf{A}} \cap \mathcal{S}$, it is sufficient to have D_λ $\bar{\alpha}$-strongly concave on $\mathcal{F}_{\mathbf{A}} \cap \mathcal{S}$ to complete the proof by following the same steps as in [12][Theorem 2]. Hence, let us show that D_λ is $\bar{\alpha}$-strongly concave in $\mathcal{F}_{\mathbf{A}} \cap \mathcal{S}$. $D_\lambda(\boldsymbol{\theta})$ is coordinate-separable with diagonal Hessian given

by $\nabla^2 D_\lambda(\boldsymbol{\theta}) = \text{Diag}\left(\left[-\lambda^2 \frac{y_i}{(1+\lambda\theta_i)^2}\right]_{i\in[m]}\right)$. The i-th eigenvalue σ_i (i.e., the i-th diagonal entry) depends only on the i-th coordinate θ_i and may tend to zero as $|\theta_i| \to +\infty$. However, when we restrict ourselves to the dual feasible set, $\mathcal{F}_\mathbf{A} = \{\boldsymbol{\theta} \mid \mathbf{A}^\mathsf{T}\boldsymbol{\theta} \leq \mathbf{1}, \boldsymbol{\theta} \geq -\mathbf{1}/\lambda\}$, $|\theta_i|$ can be bounded as follows (provided that $\text{rank}(\mathbf{A}) = \min(m,n)$)

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

The definition of $\mathcal{F}_\mathbf{A}$ implies that $-\mathbf{A}^\mathsf{T}\mathbf{1}/\lambda \leq \mathbf{A}^\mathsf{T}\boldsymbol{\theta} \leq \mathbf{1}$, or equivalently $\|\mathbf{A}^\mathsf{T}\boldsymbol{\theta}\|_\infty \leq \max\left(\|\mathbf{A}\|_1/\lambda, 1\right)$. Therefore:

$$(1+\lambda\theta_i)^2 \leq (1+\max(\|\mathbf{A}\|_1, \lambda)\|\mathbf{a}_i^\dagger\|_1)^2$$

which leads to a bound on the i-th eigenvalue. Finally, by restricting ourselves to the set $\mathcal{S} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\theta}_\mathcal{I} = -\mathbf{1}/\lambda\}$, all coordinates $i \in \mathcal{I}$ are fixed and the corresponding eigenvalues can be ignored. Hence,

$$\max_{i\in\mathcal{I}^\complement} \sigma_i \leq -\lambda^2 \min_{i\in\mathcal{I}^\complement} \frac{y_i}{(1+\max(\|\mathbf{A}\|_1, \lambda)\|\mathbf{a}_i^\dagger\|_1)^2} := -\bar{\alpha}$$

completes the proof. □

Remark 1. *To derive the safe sphere in Theorem 2, we exploit the local strong-concavity property of the dual function D_λ over $\mathcal{F}_\mathbf{A} \cap \mathcal{S}$. The restriction to the set $\mathcal{S} \subseteq \mathbb{R}^{m-|\mathcal{I}|}$ is essential to obtain this strong-concavity property as D_λ is not strongly-concave on $\mathcal{F}_\mathbf{A}$. More generally, this result is an example of how prior knowledge on the dual solution can be leveraged to improve screening strategies.*

3. PROPOSED ALGORITHM

Equipped with the screening tools derived in Section 2, we can deploy the so-called dynamic screening approach [11] to accelerate iterative solvers for problem (1). The idea is to perform screening tests repeatedly over the iterations of the underlying solver. As the algorithm converges, smaller safe regions can be defined, leading to a growing number of screened coordinates. This strategy is described in Algorithm 1 for a generic iterative solver for (1) (see Section 3.1) whose update step is denoted

$$\{\mathbf{x}, \boldsymbol{\eta}\} \leftarrow \texttt{PrimalUpdate}(\mathbf{x}, \mathbf{A}, \mathbf{y}, \lambda, \boldsymbol{\eta}). \quad (10)$$

In (10), $\mathbf{x}$ stands for the current estimate of the primal variable and $\boldsymbol{\eta}$ is a list of auxiliary variables (e.g., the gradient step-size, and possibly a few previous primal estimates). The stopping condition is given by a threshold ϵ_gap on the duality gap. Finally, in order to exploit the safe region derived in Theorem 2 for screening, it is needed to compute a dual feasible point $\boldsymbol{\theta} \in \mathcal{F}_\mathbf{A} \cap \mathcal{S}$ from the current primal estimate $\mathbf{x}$ (line 5 in Algorithm 1). This is discussed in Section 3.2.

For simplicity, in Algorithm 1 screening is performed after every iteration of the solver, but it can actually be performed at any chosen moment. For instance, it can be performed on regular intervals between a certain number of iterations of the solver. Finally, note the nested characteristic of the preserved set (line 7) as the screened coordinates are no longer tested on the ensuing iterations.

Algorithm 1 $\hat{\mathbf{x}} = \texttt{ScreeningSolverKL}(\mathbf{A}, \mathbf{y}, \lambda, \mathbf{x}, \varepsilon_\text{gap})$

1: **Initialize:** $\mathcal{A} = [n]$, $\boldsymbol{\eta}$ according to the solver
2: **repeat**
3: $\quad \{\mathbf{x}_\mathcal{A}, \boldsymbol{\eta}\} \leftarrow \texttt{PrimalUpdate}(\mathbf{x}_\mathcal{A}, \mathbf{A}_\mathcal{A}, \mathbf{y}, \lambda, \boldsymbol{\eta})$
4: $\quad$ —— *Dynamic Screening* ——
5: $\quad \boldsymbol{\theta} \leftarrow \boldsymbol{\Theta}(\mathbf{x})$ ▷ *Dual update (Proposition 2)*
6: $\quad r \leftarrow \sqrt{\frac{2\,\text{Gap}_\lambda(\mathbf{x},\boldsymbol{\theta})}{\bar{\alpha}}}$ ▷ *Safe radius (Theorem 2)*
7: $\quad \mathcal{A} \leftarrow \{j \in \mathcal{A} \mid \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta} + r\|[\mathbf{a}_j]_\mathcal{I}\|_2 \geq 1\}$
8: $\quad \mathbf{x}_{\mathcal{A}^\complement} \leftarrow \mathbf{0}$
9: **until** $\text{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta}) < \varepsilon_\text{gap}$

As screening progressively reduces the number of coordinates in play, the solver iteration cost decreases proportionally. The screening test itself does not represent a considerable overhead since it reuses the calculations already performed by the solver update. For instance, $\boldsymbol{\Theta}(\mathbf{x})$ given in Proposition 2 requires the computation of $\boldsymbol{\rho}(\mathbf{x}) = \nabla\mathcal{D}_\text{KL}(\mathbf{y} \mid \mathbf{A}\mathbf{x})$ and $\mathbf{A}^T\boldsymbol{\rho}(\mathbf{x})$, which are calculated by basically any first-order solver. Similarly, the duality gap can be reused as the stopping criterion.

3.1. Primal Update

Basically any standard iterative solver for problem (1) can be used in Algorithm 1. In the present paper, we focus on the following ones:

1. Multiplicative update (MU) [1, 2, 4, 15]
2. Proximal gradient descent (SPIRAL [7])
3. Coordinate descent (CoD) [8] (or [16])

3.2. Dual Update

The sphere center in Theorem 2 is given by a dual feasible point $\boldsymbol{\theta} \in \mathcal{F}_\mathbf{A} \cap \mathcal{S}$. Apart from primal-dual approaches such as [17, 14], most popular solvers for problem (1) only provide a primal solution estimate $\mathbf{x}$ at each iteration [4, 7, 8]. Hence, in this scenario, one needs to compute $\boldsymbol{\theta} \in \mathcal{F}_\mathbf{A} \cap \mathcal{S}$ from $\mathbf{x}$. This can be achieved at a low computational cost following Proposition 2 which is inspired from [11, 12].

Proposition 2. *Let $\mathbf{x} \in \mathbb{R}_+^n$ be a primal feasible point and set $\boldsymbol{\rho}(\mathbf{x}) = \mathbf{y}/(\mathbf{A}\mathbf{x} + \epsilon) - \mathbf{1}$. Then $\boldsymbol{\Theta} : \mathbb{R}_+^n \to \mathbb{R}_+^m$ defined by*

$$[\boldsymbol{\Theta}(\mathbf{x})]_i = \begin{cases} \frac{[\boldsymbol{\rho}(\mathbf{x})]_i}{\lambda \max(1, \max(\mathbf{A}^\mathsf{T}\boldsymbol{\rho}(\mathbf{x})/\lambda))} & \text{if } i \in \mathcal{I}^\complement \\ -\frac{1}{\lambda} & \text{if } i \in \mathcal{I} \end{cases} \quad (11)$$

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

Proof. Clearly, by definition of $\boldsymbol{\Theta}$, we have $\boldsymbol{\Theta}(\mathbf{x}) \in \mathcal{S}$. It remains to show that $\boldsymbol{\Theta}(\mathbf{x}) \in \mathcal{F}_{\mathbf{A}}$. 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 $\boldsymbol{\rho}/\lambda \geq -\mathbf{1}/\lambda$. Combining this with $\max\left(1, \max\left(\mathbf{A}^\mathsf{T}\boldsymbol{\rho}(\mathbf{x})/\lambda\right)\right) \geq 1$ and (11), we obtain that $\frac{\boldsymbol{\rho}(\mathbf{x})/\lambda}{\max(1, \max(\mathbf{A}^\mathsf{T}\boldsymbol{\rho}(\mathbf{x})/\lambda))} \geq \boldsymbol{\Theta}(\mathbf{x}) \geq -\mathbf{1}/\lambda$. Multiplying the left and right hand side of the first inequality by $\mathbf{A}^\mathsf{T}$, we obtain $\mathbf{A}^\mathsf{T}\boldsymbol{\Theta}(\mathbf{x}) \leq \mathbf{1}$ which shows that $\boldsymbol{\Theta}(\mathbf{x}) \in \mathcal{F}_{\mathbf{A}}$. Moreover, by continuity of $\boldsymbol{\rho}$ we obtain

$$\boldsymbol{\rho}(\mathbf{x}) \underset{\mathbf{x} \to \mathbf{x}^\star}{\to} \boldsymbol{\rho}(\mathbf{x}^\star) \underset{(4)}{=} \lambda \boldsymbol{\theta}^\star$$

which, together with (5) and (11) proves that $\boldsymbol{\Theta}(\mathbf{x}) \to \boldsymbol{\theta}^\star$ as $\mathbf{x} \to \mathbf{x}^\star$. □

4. EXPERIMENTS

In this section, we evaluate the performance of Algorithm 1 coupled with the three standard solvers listed in Section 3.1 for the resolution of (1) with $\epsilon = 10^{-6}$. Note that values for the hyperparameter λ are reported relatively to $\lambda_{\max} = \max\left(\mathbf{A}^\mathsf{T}(\mathbf{y} - \epsilon)\right)/\epsilon$ (the bound above which $\mathbf{x}^\star = \mathbf{0}$).

4.1. A Synthetic Toy Example

We generate synthetic data that follow the Poisson noise model $\mathbf{y} = \text{Poisson}(\mathbf{A}\mathbf{x})$ with dimensions $m = 10$, $n = 20$. The entries of $\mathbf{A}$ are drawn i.i.d. with half-normal distribution (i.e., entry $a_{ij} = |b_{ij}|$ with $b_{ij} \sim \mathcal{N}(0, 1)$), the columns $\mathbf{a}_j$ are then rescaled to unit-norm and $\mathbf{x}$ is a 2-sparse vector with uniformly-distributed nonzero entries.

Figure 1 illustrates a typical behavior of a MU solver paired with the proposed screening strategy. Screening introduces *actual* zeroes in the solution estimate, something that would not be possible with a regular MU solver, which can only shrink the coordinates without ever making them zero.

4.2. Real Datasets

We now consider real count datasets: 20-Newsgroups [18], NIPS papers [19] (word counts) and TasteProfile [20] (song listening counts). The observation vector $\mathbf{y}$ is a randomly selected column of the count data matrix and $\mathbf{A}$ is formed by the remaining data, akin to archetypal analysis.

Figure 2 shows an instance of the CoD solver's convergence over time with and without screening. Note that screening progressively reduces the iteration cost which leads to a considerable acceleration in convergence time.

Finally, Table 1 summarizes speedup results for the tested datasets with the three considered solvers in different regularization regimes and for different convergence criteria. Speedups from 1.5 to 8 times are observed, with larger gains obtained for smaller convergence tolerances (ε_{gap}).

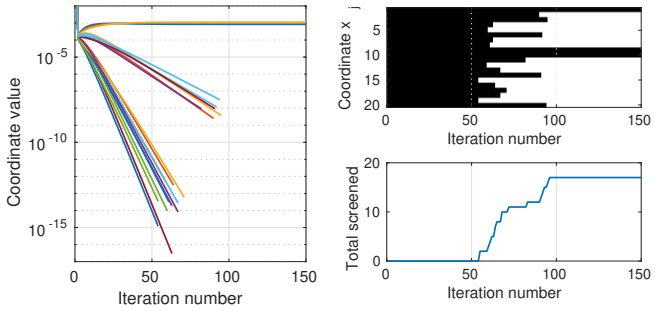

Fig. 1: Screening behavior with MU solver. Problem parameters: $m = 10$, $n = 20$, $\lambda = 10^{-3}\lambda_{\max}$. Left: each line is a coordinate x_j, the line stops when the coordinate is screened. Top: rows are coordinates which turn white when screened. Bottom: Total number of screened coordinates per iteration.

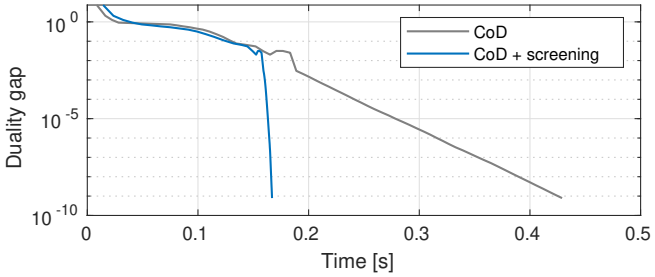

Fig. 2: Convergence of CoD solver in CPU time. 20 Newsgroups data, $\lambda = 10^{-2}\lambda_{\max}$.

5. CONCLUSION

A dynamic screening approach has been proposed to accelerate existing iterative solvers for the sparse-regularized Kullback-Leibler minimization problem. In particular, we have shown that the local strong-concavity of the associated dual function is sufficient to derive a safe screening rule. This idea could be extended for other data-fidelity terms where the dual objective function is not globally strongly concave, pushing the limits of existing safe screening rules. The proposed approach provided consistent speedups in a wide range of tested scenarios.

	$\lambda/\lambda_{\max}$	10^{-1}		10^{-3}	
	ε_{gap}	10^{-5}	10^{-7}	10^{-5}	10^{-7}
20 Newsgr.	SPIRAL	1.44	1.59	1.60	1.78
	CoD	2.44	3.42	2.46	3.22
	MU	4.80	7.72	4.49	7.28
NIPS papers	SPIRAL	2.77	3.21	2.26	2.53
	CoD	4.19	5.35	4.12	5.06
	MU	6.71	8.88	5.74	7.31
TasteProfile	SPIRAL	2.54	3.00	2.82	3.21
	CoD	1.75	2.20	2.44	4.22
	MU	2.81	4.11	2.94	4.35

Table 1: Average speedups (time without/with screening).

6. REFERENCES

[1] W. H. Richardson, "Bayesian-based iterative method of image restoration," *Journal of the Optical Society of America*, vol. 62, pp. 55–59, 1972.

[2] L. B. Lucy, "An iterative technique for the rectification of observed distributions," *Astronomical Journal*, vol. 79, pp. 745–754, 1974.

[3] V. S. Frost and B. Melamed, "Traffic modeling for telecommunications networks," *IEEE Communications Magazine*, vol. 32, no. 3, pp. 70–81, 1994.

[4] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NeurIPS)*. MIT Press, 2001.

[5] D. M. Titterington, "On the iterative image space reconstruction algorthm for ECT," *IEEE Transactions on Medical Imaging*, vol. 6, no. 1, pp. 52–56, 1987.

[6] N. Pustelnik, C. Chaux, and J. Pesquet, "Hybrid regularization for data restoration in the presence of Poisson noise," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2009.

[7] 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*, vol. 21, no. 3, pp. 1084–1096, March 2012.

[8] Cho-Jui Hsieh and Inderjit S. Dhillon, "Fast coordinate descent methods with variable selection for non-negative matrix factorization," in *Proc. International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.

[9] M. El Gheche, G. Chierchia, and J. Pesquet, "Proximity operators of discrete information divergences," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1092–1104, 2018.

[10] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani, "Safe feature elimination for the lasso and sparse supervised learning problems," *Pacific Journal of Optimization*, vol. 8, no. 4, pp. 667–698, Oct 2012, Special Issue on Conic Optimization.

[11] 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*, vol. 63, no. 19, pp. 5121–5132, Oct 2015.

[12] Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon, "Gap safe screening rules for sparsity enforcing penalties," *Journal of Machine Learning Research*, vol. 18, no. 128, pp. 1–33, Nov 2017.

[13] 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*, vol. 20, no. 163, pp. 1–42, 2019.

[14] F. Yanez and F. Bach, "Primal-dual algorithms for non-negative matrix factorization with the Kullback-Leibler divergence," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[15] Cédric Févotte and Jérôme Idier, "Algorithms for non-negative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, Sep 2011.

[16] Liangda Li, Guy Lebanon, and Haesun Park, "Fast bregman divergence nmf using taylor expansion and coordinate descent," in *Proc. International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

[17] Antonin Chambolle and Thomas Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, May 2011.

[18] "20 Newsgroups dataset," Available at: `http://qwone.com/~jason/20Newsgroups/`.

[19] "NIPS conference papers 1988-2003 dataset," Available at: `http://ai.stanford.edu/~gal/data.html`.

[20] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere, "The million song dataset," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, 2011, Available at: `http://millionsongdataset.com/tasteprofile`.