Sphere Refinement in Gap Safe Screening

Cassio F. Dantas, Emmanuel Soubies, Cédric Févotte

Sphere Refinement in Gap Safe Screening

Cássio F. Dantas, Emmanuel Soubies, Cédric Févotte, *Fellow, IEEE*

***Abstract*—The Gap safe screening technique is a powerful tool to accelerate the convergence of sparse optimization solvers. Its performance is largely based on the ability to determine the smallest "sphere", centered at a given feasible dual point, that contains the dual solution. This can be achieved through an inner sphere refinement loop, applied at each screening step. In this work, we show that this refinement loop actually converges to the solution of a fixed-point equation for which we derive a closed-form expression for two common loss functions. This allows us to develop an analytic (i.e., non iterative), more concise and theoretically-grounded variant of the sphere refinement step.**

***Index Terms*—Sparse optimization, Safe screening, Kullback-Leibler regression, Logistic regression.**

I. Introduction

SPARSE optimization problems are encountered in fields such as signal processing, inverse problems, statistics, and machine learning. A very common formulation is given by

$$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{C}} P_\lambda(\mathbf{x}) := \sum_{i=1}^{m} f_i([\mathbf{A}\mathbf{x}]_i) + \lambda\|\mathbf{x}\|_1 \qquad (1)$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathcal{C} \in \{\mathbb{R}^n, \mathbb{R}^n_{\geq 0}\}$, $\lambda > 0$, and each scalar function $f_i : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is proper convex, and differentiable. As such, numerous algorithms have been developed to tackle problems of the form (1). These include, but are not limited to, proximal gradient [1]–[3], coordinate descent [4], [5], and majorization-minimization [6] methods.

Within this context, the promise of safe screening is to identify zero coordinates in $\mathbf{x}^\star$ so as to reduce the size of the problem and, consequently, accelerate the convergence of the solver. This identification can be performed before or within the course of iterations, leading respectively to the so-called static [7] and dynamic [8] screening approaches. Although originally proposed for the Lasso problem [7] (i.e., $f_i(z) = (y_i - z)^2$ where y_i is the ith entry of a data vector $\mathbf{y} \in \mathbb{R}^m$), safe screening techniques have then been extended to a large variety of sparse-regularized problems [8]–[11]. The case where the ℓ_1-norm in (1) is replaced by a generic group separable norm has also been treated in [9], [12].

Notations: We let $[n] = \{1, \dots, n\}$. For a vector $\mathbf{x} \in \mathbb{R}^n$, we denote x_i its ith entry. Given a subset of indices $g \subseteq [n]$ with cardinality $|g| = n_g$, $\mathbf{x}_g \in \mathbb{R}^{n_g}$ is the restriction of $\mathbf{x}$ to its elements indexed by g. For a matrix $\mathbf{A}$, we similarly define the restrictions $\mathbf{a}_j$ and $\mathbf{A}_g$ with respect to the columns of $\mathbf{A}$. The complement of $\mathcal{A} \subseteq [n]$ is denoted $\mathcal{A}^c = [n] \setminus \mathcal{A}$.

This work was supported by the European Research Council (FACTORY, ERC-CoG-6681839), the French ANR (EROSION, ANR-22-CE48-0004) and the NRF in Singapore (DesCartes, CREATE program).

CFD is with TETIS, Université Montpellier, INRAE, Montpellier, France (email: cassio.fraga-dantas@inrae.fr). ES and CF are with IRIT, Université de Toulouse, CNRS, Toulouse, France (e-mail: firstname.lastname@irit.fr).

Safe Screening in a Nutshell: Safe screening techniques rely on the first-order primal-dual optimality conditions of (1). More precisely, we get the key property that [7], [12]

$$\forall j \in [n],\ |\phi(\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}^\star)| < 1 \implies x_j^\star = 0, \qquad (2)$$

where $\phi(x) = x$ if $\mathcal{C} = \mathbb{R}^n$ or $\phi(x) = \max(x, 0)$ if $\mathcal{C} = \mathbb{R}^n_{\geq 0}$, and $\boldsymbol{\theta}^* \in \mathbb{R}^m$ is the solution of the dual problem

$$\boldsymbol{\theta}^\star = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Delta_\mathbf{A}} D_\lambda(\boldsymbol{\theta}) := -\textstyle\sum_{i=1}^m f_i^*(-\lambda\theta_i). \qquad (3)$$

In the dual formulation (3), f_i^* stands for the Fenchel-Legendre conjugate of f_i while $\Delta_\mathbf{A} := \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \forall j \in [n],\ |\phi(\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta})| \leq 1\} \cap \operatorname{dom}(D_\lambda)$ corresponds to the dual feasible set, with $\operatorname{dom}(D_\lambda)$ denoting the domain of the dual function.

One sees from (2) that the knowledge of $\boldsymbol{\theta}^\star$ allows us to identify zero coordinates in $\mathbf{x}^\star$. Yet, this is not practical as $\boldsymbol{\theta}^\star$ is unknown. The main task in safe screening is thus to define a *safe region* $\mathcal{S} \ni \boldsymbol{\theta}^\star$ from which we can derive the following *safe screening rule* for the jth component

$$\max_{\boldsymbol{\theta}\in\mathcal{S}} |\phi(\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta})| < 1 \implies |\phi(\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}^\star)| < 1 \implies x_j^\star = 0. \qquad (4)$$

Clearly, in order to maximize screening performance, the safe region $\mathcal{S}$ should be as small as possible (to increase the number of screened variables) while allowing an efficient computation of the *screening test* given by $\max_{\boldsymbol{\theta}\in\mathcal{S}} |\phi(\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta})| < 1$ (to minimize the computational overhead).

Among existing safe regions, the Gap safe sphere [9] leads to state-of-the-art screening performance. Given any primal-dual pair $(\mathbf{x}, \boldsymbol{\theta})$, it reads as $\mathcal{S} = \mathcal{B}(\boldsymbol{\theta}, \sqrt{2\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})/\alpha_{\mathbb{R}^m}})$, where $\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta}) = P_\lambda(\mathbf{x}) - D_\lambda(\boldsymbol{\theta}) \geq 0$ and $\alpha_{\mathbb{R}^m} > 0$ corresponds to the strong concavity constant of D_λ over $\mathbb{R}^m$. Not only is its geometry simple (allowing fast screening tests), but its radius vanishes upon convergence of the primal-dual iterates when strong duality holds (i.e., $\mathrm{Gap}_\lambda(\mathbf{x}^\star, \boldsymbol{\theta}^\star) = 0$). Yet, it requires the dual function D_λ to be globally strongly concave which precludes its use for an important class of functions f_i such as the β-divergences with $\beta \in [1, 2)$ [13].

In a previous work [12], we overcame this limitation by computing local strong concavity bounds on well-chosen subsets of the domain. Moreover, by re-evaluating the strong-concavity bound on the current safe sphere, we proposed a sphere refinement loop that improves screening performance.

Contributions: In this letter, we scrutinize the sphere refinement loop proposed in [12] and recalled in Sec. II. We prove that it converges to the solution of a fixed-point equation (Proposition 1). This allows us to derive a new algorithm that is exempt of the inner loop, when the fixed-point equation admits a closed-form solution (Alg. 2). We derive in Sec. IV such closed-from expressions for two popular loss functions: the Kullback-Leibler (KL) divergence and the logistic function. Numerical evaluations are reported in Sec. V. Finally, all proof are deferred to the *Supplementary Material*.

Algorithm 1 GSS with iterative sphere refinement [12]

1: **Inputs:** $\mathbf{x}^0 \in \mathcal{C}$, $\varepsilon_{\text{gap}} > 0$
2: $\mathcal{A} \leftarrow [n]$, $\mathcal{S}^0 \leftarrow \Delta_{\mathbf{A}}$, $\varepsilon_r \leftarrow 10^{-3}$
3: **repeat** *(loop over k)*
4: ▷ *Primal and Dual updates*
5: $\mathbf{x}^k_{\mathcal{A}} \leftarrow \texttt{PrimalUpdate}(\mathbf{x}^{k-1}_{\mathcal{A}}, \mathbf{A}_{\mathcal{A}})$, $\mathbf{x}^k_{\mathcal{A}^c} \leftarrow \mathbf{0}$
6: $\boldsymbol{\theta}^k \leftarrow \texttt{DualUpdate}(\mathbf{x}^k) \in \Delta_{\mathbf{A}}$
7: ▷ *Safe region with iterative refinement*
8: $\tilde{\mathcal{S}} \leftarrow \mathcal{S}^{k-1}$
9: **if** $\boldsymbol{\theta}^k \notin \tilde{\mathcal{S}}$ **then** ▷ *Inflate previous safe region*
10: $\tilde{\mathcal{S}} \leftarrow \mathcal{B}\left(\boldsymbol{\theta}^{k-1}, \|\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}\|\right)$
11: **end if**
12: $\mathcal{S}^k \leftarrow \mathcal{B}\left(\boldsymbol{\theta}^k, \sqrt{\frac{2\,\text{Gap}_\lambda(\mathbf{x}^k, \boldsymbol{\theta}^k)}{\alpha_{\tilde{\mathcal{S}}}}}\right)$ ▷ *Init. safe region*
13: **if** $\alpha_{\mathcal{S}^k} > \alpha_{\tilde{\mathcal{S}}}$ **then**
14: $\alpha_0 \leftarrow \alpha_{\mathcal{S}^k}$
15: **repeat** *(loop over j)* ▷ *Sphere refinement*
16: $r_j \leftarrow \sqrt{2\,\text{Gap}_\lambda(\mathbf{x}^k, \boldsymbol{\theta}^k)/\alpha_{j-1}}$
17: $\alpha_j \leftarrow \alpha_{\mathcal{B}(\boldsymbol{\theta}^k, r_j)}$
18: **until** $|r_j - r_{j-1}| < \varepsilon_r \cdot r_{j-1}$
19: $\mathcal{S}^k \leftarrow \mathcal{B}(\boldsymbol{\theta}^k, r_j)$
20: **end if**
21: ▷ *Screening*
22: $\mathcal{A} \leftarrow \{j \in \mathcal{A} | \max_{\boldsymbol{\theta} \in \mathcal{S}^k} |\phi(\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta})| \geq 1\}$
23: **until** $\text{Gap}_\lambda(\mathbf{x}^k, \boldsymbol{\theta}^k) \leq \varepsilon_{\text{gap}}$

II. GSS with Iterative Sphere Refinement

The Gap safe screening (GSS) with iterative sphere refinement proposed in [12] is recalled in Alg. 1. There, `PrimalUpdate` (resp., `DualUpdate`) refers to the update step of any iterative primal (resp., dual) solver for (1) (resp., (3)). Given a subset $\mathcal{S} \subset \mathbb{R}^m$, $\alpha_{\mathcal{S}}$ stands for the strong concavity constant of D_λ over $\mathcal{S}$, i.e., a lower bound on $f_i^{*\prime\prime}$ over $\mathcal{S}$ and $\forall i \in [m]$ (c.f. [14], [12, Proposition 11]). Then, the construction of the safe region, starting at line 7, is made of three steps.

- First, if the new dual point does not belong to the previous safe region, the latter is inflated (Line 10) to make the next step possible with $\boldsymbol{\theta}^k \in \tilde{\mathcal{S}}$.
- Second, as $\boldsymbol{\theta}^k \in \tilde{\mathcal{S}}$, Theorem 5 in [12] can be invoked to build a new safe region centered at $\boldsymbol{\theta}^k$ using $\alpha_{\tilde{\mathcal{S}}}$ (Line 12).
- Third, if the strong concavity constant over this new safe region improves (Line 13), this safe region is iteratively refined (lines 15–18). From [12, Proposition 7], this refinement loop generates a sequence of nested Gap safe spheres (i.e., with decreasing radius), all centered in $\boldsymbol{\theta}^k$.

Finally, the refined safe region is used at Line 22 to safely screen out zero-coordinates of the solution vector $\mathbf{x}^\star$.

III. GSS with Analytic Sphere Refinement

The proposed Gap safe screening with analytic sphere refinement is presented in Alg. 2. Its main novelties with respect to Alg. 1 are outlined in the next three sections.

A. Tracking the Best Strong Concavity Constant

As opposed to Alg. 1, in Alg. 2 we keep track of the safe region $\mathcal{S}_{\text{b}}$ over which the best (i.e., largest) constant $\alpha_{\mathcal{S}_{\text{b}}}$ has been computed so far. As such, we ensure the construction of a non-decreasing sequence of strong concavity constants. Then, to ensure that the new dual point $\boldsymbol{\theta}^k$ belongs to $\mathcal{S}_{\text{b}}$, we replaced the inflation step at Line 10 of Alg. 1 by the projection step at Line 7 of Alg. 2. The benefit of this modification is twofold.

- It discards the need of recomputing the strong concavity constant (on the inflated region) before refinement.
- It leads to improved dual points. Indeed, given that $\mathcal{S}_{\text{b}}$ is convex and $\boldsymbol{\theta}^\star \in \mathcal{S}_{\text{b}}$, we have for all $\tilde{\boldsymbol{\theta}} \in \Delta_{\mathbf{A}}$ $\|P_{\mathcal{S}_{\text{b}}}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{\theta}^\star\|_2 \leq \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2$.

Algorithm 2 Proposed GSS with analytic sphere refinement

1: **Inputs:** $\mathbf{x}^0 \in \mathcal{C}$, $\varepsilon_{\text{gap}} > 0$
2: $\mathcal{A} \leftarrow [n]$, $\mathcal{S}_{\text{b}} \leftarrow \Delta_{\mathbf{A}}$
3: **repeat** *(loop over k)*
4: ▷ *Primal and Dual updates*
5: $\mathbf{x}^k_{\mathcal{A}} \leftarrow \texttt{PrimalUpdate}(\mathbf{x}^{k-1}_{\mathcal{A}}, \mathbf{A}_{\mathcal{A}})$, $\mathbf{x}^k_{\mathcal{A}^c} \leftarrow \mathbf{0}$
6: $\tilde{\boldsymbol{\theta}} \leftarrow \texttt{DualUpdate}(\mathbf{x}^k) \in \Delta_{\mathbf{A}}$
7: $\boldsymbol{\theta}^k \leftarrow P_{\mathcal{S}_{\text{b}}}(\tilde{\boldsymbol{\theta}})$ ▷ *Projection onto $\mathcal{S}_{\text{b}}$ (c.f. [14])*
8: ▷ *Safe region with analytic refinement*
9: $\mathcal{S}^k \leftarrow \mathcal{B}\left(\boldsymbol{\theta}^k, \sqrt{\frac{2\,\text{Gap}_\lambda(\mathbf{x}^k, \boldsymbol{\theta}^k)}{\alpha_{\mathcal{S}_{\text{b}}}}}\right)$ ▷ *Init. safe region*
10: **if** $\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_{\mathcal{S}_{\text{b}}}\| > r_{\mathcal{S}^k} - r_{\mathcal{S}_{\text{b}}}$ **then**
11: **if** $\bar{\alpha}^k > \alpha_{\mathcal{S}_{\text{b}}}$ **then** ▷ *$\bar{\alpha}^k$ fixed-point of (8)*
12: $\mathcal{S}^k \leftarrow \mathcal{B}\left(\boldsymbol{\theta}^k, \sqrt{\frac{2\,\text{Gap}_\lambda(\mathbf{x}^k, \boldsymbol{\theta}^k)}{\bar{\alpha}^k}}\right)$
13: $\mathcal{S}_{\text{b}} \leftarrow \mathcal{S}^k$ ▷ *Track region with best α*
14: **end if**
15: **end if**
16: ▷ *Screening*
17: $\mathcal{A} \leftarrow \{j \in \mathcal{A} | \max_{\boldsymbol{\theta} \in \mathcal{S}^k} |\phi(\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta})| \geq 1\}$
18: **until** $\text{Gap}_\lambda(\mathbf{x}^k, \boldsymbol{\theta}^k) \leq \varepsilon_{\text{gap}}$

B. Avoiding Unnecessary Refinement Attempts

In Alg. 2, we added the test at Line 10 to avoid unnecessary refinement attempts. Indeed, one can see that the refinement step will improve the initial kth Gap safe sphere

$$\mathcal{S}^k = \mathcal{B}\left(\boldsymbol{\theta}^k, r^k\right) \quad \text{with} \quad r^k = \sqrt{\frac{2\,\text{Gap}_\lambda(\mathbf{x}^k, \boldsymbol{\theta}^k)}{\alpha_{\mathcal{S}_{\text{b}}}}}, \quad (5)$$

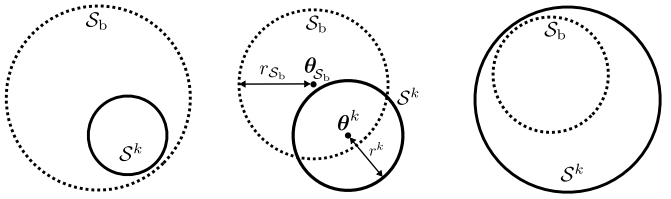
only if $\alpha_{\mathcal{S}^k} > \alpha_{\mathcal{S}_{\text{b}}}$ (i.e., the strong concavity constant on the new $\mathcal{S}^k$ is better than the best one $\alpha_{\mathcal{S}_{\text{b}}}$ computed so far). From the definition of strong concavity,[1] we can thus derive the following three situations (illustrated in Fig. 1 (a-c)),

- *Improvement* if $\mathcal{S}^k \subseteq \mathcal{S}_{\text{b}} \Longrightarrow \alpha_{\mathcal{S}^k} \geq \alpha_{\mathcal{S}_{\text{b}}}$
- *No Improvement* if $\mathcal{S}^k \supseteq \mathcal{S}_{\text{b}} \Longrightarrow \alpha_{\mathcal{S}^k} \leq \alpha_{\mathcal{S}_{\text{b}}}$
- *Indecisive* otherwise.

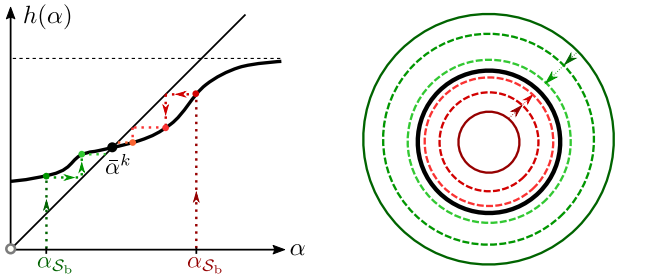
A typical situation of *improvement* arises when the duality Gap (and thus the radius) decreases more than the displacement of the dual point from one iteration to the next. Indeed, let $\boldsymbol{\theta}_{\mathcal{S}_{\text{b}}} \in \Delta_{\mathbf{A}}$ and $r_{\mathcal{S}_{\text{b}}} > 0$ denote respectively the center and the radius of $\mathcal{S}_{\text{b}}$, then $\mathcal{S}^k \subseteq \mathcal{S}_{\text{b}}$ is equivalent to (see Fig. 1-a)

$$\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_{\mathcal{S}_{\text{b}}}\| \leq r_{\mathcal{S}_{\text{b}}} - r^k. \quad (6)$$

[1] If a function f is α_1-strongly concave on $\mathcal{S}_1$, then it is $\alpha_2 \geq \alpha_1$ strongly concave on any $\mathcal{S}_2 \subset \mathcal{S}_1$.



(a) Improvement (b) Indecisive (c) No Improvement



(d) Refinement loop (α) (e) Refinement loop (sphere)

Fig. 1: Sphere refinement behavior. In case (a), computing the fixed point $\bar{\alpha}^k$ would improve over $\alpha_{\mathcal{S}_\mathrm{b}}$. The refinement loop would follow the green path in (d-e). On the contrary, in case (c), computing the fixed point $\bar{\alpha}^k$ would degrade w.r.t. $\alpha_{\mathcal{S}_\mathrm{b}}$. There, running the refinement loop would follow the red path in (d-e). This situation is avoided in Alg. 1 (resp. Alg. 2) by the test at Line 13 (resp. Lines 10-11). Finally, the intermediate case (b) may lead to both mentioned behaviors.

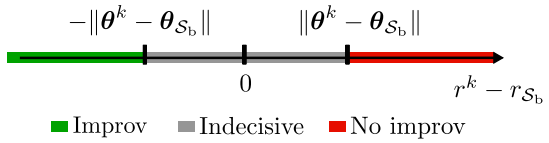


Fig. 2: Illustration of the test at Line 10 of Alg. 2

Similarly, we get that $\mathcal{S}^k \supseteq \mathcal{S}_\mathrm{b}$ (i.e., *no improvement* case) is equivalent to

$$\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_{\mathcal{S}_\mathrm{b}}\| \leq r^k - r_{\mathcal{S}_\mathrm{b}}. \tag{7}$$

Hence, the complement of (7) includes all *improvement* and *indecisive* cases (see Fig. 2). It can be used as a test that does not require to compute any strong concavity constant to decide whether to perform the refinement step (Line 10 of Alg. 2). Yet, a second test involving strong concavity constants (Line 11 of Alg. 2) is required to deal with *indecisive* cases.

C. Dropping the Refinement Loop

In Proposition 1, we prove that the sequence of strong concavity constants generated by the refinement loop at lines 15–18 of Alg. 1 converges to the solution $\bar{\alpha}^k$ of a fixed-point equation. This is illustrated in Fig. 1 (d-e). As such, provided that one has access to a closed-form expression for $\bar{\alpha}^k$ (see Sec. IV), the refinement is no longer iterative, as implemented at Line 12 of Alg. 2.

Proposition 1 (Fixed point equation)**.** *Assume that D_λ is twice differentiable and let $(\mathbf{x}^k, \boldsymbol{\theta}^k) \in \mathcal{C} \times \Delta_\mathbf{A}$ be the kth primal-dual iterate pair and $G^k := \mathrm{Gap}_\lambda(\mathbf{x}^k, \boldsymbol{\theta}^k)$. Define*

$$h^k(\alpha) = \min_{i \in [m]} \inf_{\theta_i \in \mathcal{B}(\theta_i^k, \sqrt{\frac{2G^k}{\alpha}}) \cap \mathrm{dom}\, f_i^*(-\lambda \cdot)} \lambda^2 f_i^{*\prime\prime}(-\lambda\theta_i) \tag{8}$$

with unconstrained inf *when $\alpha = 0$. Then, the refinement loop (lines 15–18 in Alg. 1) converges to the safe region $\mathcal{S}^k = \mathcal{B}\left(\boldsymbol{\theta}^k, \sqrt{2G^k/\bar{\alpha}^k}\right)$ with $\bar{\alpha}^k$ a non-repelling[2] fixed point of h^k.*

IV. Closed-Form Expressions of Fixed Points

The practical relevance of Alg. 2 depends on our ability to derive closed-form expressions of fixed points of h^k in (8). We show that this is possible for two very common loss functions. More general properties of h^k are given in [14, Sec. 1.3].

A. Kullback-Leibler Divergence

Here, the scalar data-fidelity functions f_i and their convex conjugates are given by:

$$f_i(z) = y_i \log\left(y_i/(z+\epsilon)\right) + z + \epsilon - y_i, \tag{9}$$
$$f_i^*(u) = -y_i \log(1-u) - \epsilon u, \tag{10}$$

where y_i is the ith entry of the data vector $\mathbf{y} \in \mathbb{R}^m_{\geq 0}$, $\epsilon > 0$ is a smoothing factor that avoids singularities around zero and $\mathrm{dom}(f_i^*) = \{u \in \mathbb{R} \mid u \leq 1\}$. Finally, in this case we have $\mathcal{C} = \mathbb{R}^n_{\geq 0}$, and $\mathbf{A} \in \mathbb{R}^{m \times n}_{\geq 0}$.

Proposition 2. *Assume that $y_i > 0$ for all $i \in [m]$. Let $(\mathbf{x}^k, \boldsymbol{\theta}^k) \in \mathbb{R}^n_{\geq 0} \times \Delta_\mathbf{A}$ and $G^k := \mathrm{Gap}_\lambda(\mathbf{x}^k, \boldsymbol{\theta}^k)$. Then, h^k in (8) with f_i^* in (10) has a* unique *attracting fixed-point*

$$\bar{\alpha}^k = \min_{i \in [m]} \bar{\alpha}_i^k \tag{11}$$

$$\text{with } \bar{\alpha}_i^k = \begin{cases} 0 & \text{if } G^k \geq \dfrac{y_i}{2} \\ \dfrac{\lambda^2 \left(\sqrt{y_i} - \sqrt{2G^k}\right)^2}{(1+\lambda\theta_i^k)^2} & \text{otherwise.} \end{cases} \tag{12}$$

Remark 1. *Following [12, Sec. 4.2.2], Proposition 2 can be generalized to the situation where there exists $i \in [m]$ such that $y_i = 0$. This is achieved by searching for the* min *in* (11) *within $\mathcal{I}_0^\complement$ rather than $[m]$, where $\mathcal{I}_0 = \{i \in [m] \mid y_i = 0\}$.*

B. Logistic Function

For an input signal $\mathbf{y} \in \mathbb{R}^m$, the data-fidelity functions f_i and their convex conjugates f_i^* are given by:

$$f_i(z) = \log\left(1 + e^z\right) - y_i z \tag{13}$$
$$f_i^*(u) = (y_i+u)\log(y_i+u) + (1-y_i-u)\log(1-y_i-u)$$

with $\mathrm{dom}(f_i^*) = \{u \in \mathbb{R} \mid 0 \leq u + y_i \leq 1\} = [-y_i, 1-y_i]$. In this case, we have $\mathcal{C} = \mathbb{R}^n$.

Proposition 3. *Let $(\mathbf{x}^k, \boldsymbol{\theta}^k) \in \mathbb{R}^n \times \Delta_\mathbf{A}$ and $G^k := \mathrm{Gap}_\lambda(\mathbf{x}^k, \boldsymbol{\theta}^k)$. Then h^k in (8) with f_i^* in (13) has a* unique

[2] See [15, Definitions 1.3 and 1.4] or [16, Definitions 6.5 and 6.8] for a definition of attracting and repelling fixed points.

TABLE I: Proportion of times (%) where $\bar{\alpha}^k \leq \alpha_{\mathcal{S}_\text{b}}$ or $\bar{\alpha}^k > \alpha_{\mathcal{S}_\text{b}}$ for each case of Fig. 1. Results are averaged over a regularization grid of 100 points $\lambda/\lambda_\text{max} \in [10^{-3}, 1)$ (standard deviation is shown in parentheses). Bold numbers emphasize the most critical case where $\bar{\alpha}^k$ is computed but not used.

	CoD for Logistic		Prox. Grad. for KL	
	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_\text{b}}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_\text{b}}$	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_\text{b}}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_\text{b}}$
Improv.	0	92.2 (5.3)	0	45.6 (6.1)
No-Improv.	5.9 (4.6)	0	53.2 (6.4)	0
Indec.	**1.0** (0.7)	0.9 (0.7)	**0.6** (1.1)	0.6 (1.2)

attracting fixed-point $\bar{\alpha}^k = \min_{i\in[m]} \bar{\alpha}_i^k$ *with, for* $\tau_i = |\lambda\theta_i^k - y_i + \frac{1}{2}| \leq \frac{1}{2}$,

$$\bar{\alpha}_i^k = \begin{cases} 4\lambda^2 & \text{if } G^k \geq 2\tau_i^2 \\ \dfrac{\lambda^2(2G^k+1)^2}{2G^k} & \text{if } G^k < 2\tau_i^2 \\ & \text{and } \tau_i = \frac{1}{2} \\ \left(\dfrac{-4\tau_i\lambda\sqrt{2G^k}+2\lambda\sqrt{2G^k+1-4\tau_i^2}}{1-4\tau_i^2}\right)^2 & \text{if } G^k < 2\tau_i^2 \\ & \text{and } \tau_i < \frac{1}{2} \end{cases} \quad (14)$$

C. Complexity Analysis

The computation of the fixed point from the closed-form expressions derived in propositions 2 and 3 is of the order of $O(m)$ (due to the min operations). This is about the same complexity as for the evaluation of the strong concavity constant on any ball [12, Table 1]. As such, denoting by P the number of sphere refinement iterations in Alg. 1, the analytic version of Alg. 2 allows for reducing the overall sphere refinement complexity from $O(Pm)$ to $O(m)$.

V. Numerical Illustration

In this section, we illustrate the behavior of the sphere refinement procedure with the following two examples:

- KL regression with a proximal gradient (PG) solver [3] for archetypal analysis on the NIPS papers dataset [17]. The size of the problem is $(m \times n) = (2483 \times 14035)$ and $\lambda/\lambda_\text{max} = 10^{-1}$ (λ_max being the regularization above which the zero vector is a solution of (1) [12, Sec. 4.1.2]).
- Logistic regression with a coordinate descent (CoD) solver [18] for binary classification of the Leukemia dataset [19]. The size of the problem is $(m \times n) = (71 \times 7129)$ and $\lambda/\lambda_\text{max} = 10^{-2}$.

We report in Table I, for each of the three situations described in Fig. 1, the distribution of times where $\bar{\alpha}^k > \alpha_{\mathcal{S}_\text{b}}$ and $\bar{\alpha}^k \leq \alpha_{\mathcal{S}_\text{b}}$. The reported distributions have been experimentally observed to depend mostly on the underlying solver rather than other parameters like regularization and problem instance. Therefore, the reported scenarios in Table I are quite representative of CoD and PG solver typical behaviors [14]. An interesting observation is that, for both experiments, the proportion of *indecisive* situations is very low (1.9% for Logistic regression with CoD and 1.2% for KL-regression with proximal gradient). In particular, the proportion of times where the fixed point $\bar{\alpha}^k$ has been computed without being used

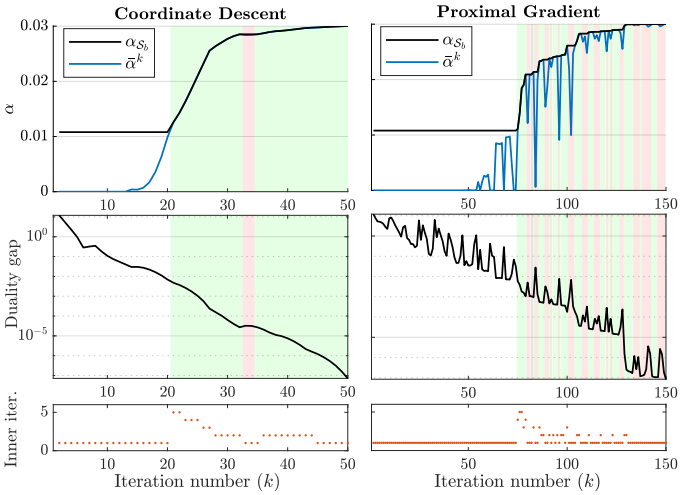


Fig. 3: Evolution of $\bar{\alpha}^k$ and $\alpha_{\mathcal{S}_\text{b}}$ (top), the duality gap G^k (middle), and the number of inner refinement iterations performed by Alg. 1 (bottom), as a function of the outer iteration number k. Green, grey (very rare) and red backgrounds depict respectively the *improvement*, *indecisive*, and *no-improvement* situations. The initial white region corresponds to a "burn-in" phase where $\bar{\alpha}^k < \alpha_{\mathcal{S}_\text{b}} = \alpha_{\Delta_\mathbf{A}}$. Indeed, here $\alpha_{\mathcal{S}_\text{b}}$ has been initialized with $\alpha_{\Delta_\mathbf{A}}$ which is known.

(bold values in Table I) is even smaller. These observations show the efficiency of the simple test at Line 10 of Alg. 2 in discriminating *improvement* from *no-improvement* situations. If, for a given problem, *indecisive* cases are more abundant, often leading to useless computations of $\bar{\alpha}^k$, one can modify the test in Line 10 so as to exclude such *indecisive* cases.

To further illustrate the sphere refinement behavior, we report in Fig. 3 the evolution of the best strong concavity constant $\alpha_{\mathcal{S}_\text{b}}$, the kth fixed point $\bar{\alpha}^k$, and the duality gap $G^k = \text{Gap}_\lambda(\mathbf{x}^k, \boldsymbol{\theta}^k)$ as a function of the iteration number k. We observe that, in general, *no-improvement* situations (red areas) occur at iterations where the duality gap increased. This behavior can be more frequent for some solvers (e.g., PG) than others (e.g., CoD). Moreover, we see that the fixed point $\bar{\alpha}^k$ (blue curves) can be significantly degraded in such *no-improvement* situations. This shows the importance of updating the strong concavity constant only when it improves over the current one ($\alpha_{\mathcal{S}_\text{b}}$).

The number of refinement iterations performed by Alg. 1 is also reported in Fig. 3. We see that, in the coordinate descent (resp. proximal gradient) case, a total of 37 (resp. 69) inner iterations are avoided by the proposed approach (more extensive experiments are available in [14]).

VI. Conclusion

In this work, we made a detailed theoretical analysis of the sphere refinement loop proposed in [12], by reformulating it as fixed point iterations and characterizing its convergence point. Not only does it shed new light on this refinement step, but it allows us to derive a non-iterative version that is more elegant, more concise, and enjoys a better computational complexity.

REFERENCES

[1] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005. [Online]. Available: https://doi.org/10.1137/050626090

[2] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, Jan 2009. [Online]. Available: http://dx.doi.org/10.1137/080716542

[3] 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.

[4] W. J. Fu, "Penalized regressions: The bridge versus the lasso," *Journal of Computational and Graphical Statistics*, vol. 7, no. 3, pp. 397–416, 1998. [Online]. Available: http://www.jstor.org/stable/1390712

[5] G.-X. Yuan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin, "A comparison of optimization methods and software for large-scale l1-regularized linear classification," *Journal of Machine Learning Research*, vol. 11, no. 105, pp. 3183–3234, 2010. [Online]. Available: http://jmlr.org/papers/v11/yuan10c.html

[6] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization–minimization algorithms for wavelet-based image restoration," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2980–2991, 2007.

[7] L. El Ghaoui, V. Viallon, and T. 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.

[8] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. 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.

[9] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon, "Gap safe screening rules for sparsity enforcing penalties," *Journal of Machine Learning Research*, vol. 18, no. 128, pp. 1–33, Nov 2017.

[10] J. Wang, Z. Zhang, and J. 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. [Online]. Available: http://jmlr.org/papers/v20/16-383.html

[11] C. F. Dantas, E. Soubies, and C. 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.

[12] ——, "Expanding boundaries of Gap Safe screening," *Journal of Machine Learning Research (JMLR)*, vol. 22, no. 236, pp. 1–57, 2021.

[13] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, p. 2421–2456, Sep 2011. [Online]. Available: https://doi.org/10.1162/NECO_a_00168

[14] C. F. Dantas, E. Soubies, and C. Févotte, "Supplementary Material to Sphere Refinement in Gap Safe Screening," available at: https://hal.science/hal-03891840.

[15] A. Johnson, K. Madden, and A. Sahin, *Discovering Discrete Dynamical Systems*. American Mathematical Soc., 2017, vol. 53.

[16] A. Granas and J. Dugundji, *Fixed point theory*. Springer, 2003, vol. 14.

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

[18] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Mathematical Programming*, vol. 117, no. 1, pp. 387–423, 2009.

[19] 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*, vol. 286, pp. 531–537, 1999.

SUPPLEMENTARY MATERIAL

A. Proof sketch of Proposition 1

From the definition of h^k in (8), we get that the refinement loop at lines 15–18 of Alg. 1 amounts to the fixed-point iteration $\alpha_j = h^k(\alpha_{j-1})$. Moreover, we know from [12, Proposition 7] that this refinement loop builds a sequence of nested Gap Safe spheres (i.e., with decreasing radius). Hence, the sequence $(\alpha_j)_j$ converges and its limit point $\bar{\alpha}^k$ is a fixed-point of h^k. Moreover, it is non-repelling as it can be reached with the fixed-point iteration at least from one initial point. A detailed proof of Proposition 1 is provided in [14, Sec. 1.2].

B. Proof of Propositions 2 and 3

Lemma 1. *[15, Theorem 2.1] Let $h \in C^1$ having a fixed point at $\bar{\alpha}$. Then $\bar{\alpha}$ is attracting if $|h'(\bar{\alpha})| < 1$ and repelling if $|h'(\bar{\alpha})| > 1$.*

Lemma 2. *For $i \in [m]$, let $h_i : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a continuous function that has i) a unique attracting fixed point $\bar{\alpha}_i \geq 0$ and ii) potentially 0 as repelling fixed point. Then $\bar{\alpha} = \min_{i\in[m]} \bar{\alpha}_i$ is the unique attracting fixed point of $h : \alpha \mapsto \min_{i\in[m]} h_i(\alpha)$.*

Proof. By assumption on the h_i, we have $h_i(\bar{\alpha}_i) = \bar{\alpha}_i$ and

$$h_i(\alpha) < \alpha,\ \forall \alpha \in (\bar{\alpha}_i, +\infty) \text{ and } h_i(\alpha) > \alpha,\ \forall \alpha \in (0, \bar{\alpha}_i).$$

Denoting $i^\star = \operatorname{argmin}_{i\in[m]} \bar{\alpha}_i$, we get

$$h(\alpha) \leq h_{i^\star}(\alpha) < \alpha,\ \forall \alpha \in (\bar{\alpha}_{i^\star}, +\infty) \tag{15}$$

$$h(\alpha) > \alpha,\ \forall \alpha \in (0, \bar{\alpha}_{i^\star}), \tag{16}$$

The continuity of h_i and h completes the proof. □

Define h_i^k such that h^k in (8) can be written as

$$h^k(\alpha) = \min_{i\in[m]} h_i^k(\alpha). \tag{17}$$

Then, to prove Propositions 2 and 3, it suffices to show that the corresponding h_i^k fulfill the conditions of Lemma 2.

Proof of Proposition 2

We get from [12, Proposition 36] (with $r = \sqrt{2G^k}/\alpha$) that

$$h_i^k(\alpha) := \frac{\lambda^2 y_i \alpha}{\left(\sqrt{\alpha}(1+\lambda\theta_i^k) + \lambda\sqrt{2G^k}\right)^2},$$

$$(h_i^k)'(\alpha) := \frac{\lambda^3 y_i \sqrt{2G^k}}{\left(\sqrt{\alpha}(1+\lambda\theta_i^k) + \lambda\sqrt{2G^k}\right)^3}.$$

In order to invoke Lemma 2, let us analyze the fixed point of h_i^k. Clearly, we always have $h_i^k(0) = 0$ showing that 0 is a fixed point of h_i^k. Moreover, from Lemma 1, it is attracting iff $(h_i^k)'(0) = y_i/(2G^k) < 1 \Leftrightarrow G^k > y_i/2$. Concerning non-zero fixed points, they satisfy

$$\alpha(1+\lambda\theta_i^k)^2 + 2\sqrt{\alpha}(1+\lambda\theta_i^k)\lambda\sqrt{2G^k} + \lambda^2(2G^k - y_i) = 0.$$

This is a quadratic equation in $\sqrt{\alpha}$ with solutions

$$\sqrt{\alpha} = \frac{\lambda(-\sqrt{2G^k} \pm \sqrt{y_i})}{1+\lambda\theta_i^k}. \tag{18}$$

Because $-\sqrt{2G^k} \leq 0$, only the "plus" solution is admissible when $G^k < y_i/2$. Moreover, we get from Lemma 1 that $G^k < y_i/2$ also implies that this non-zero fixed point is attracting.

Combining the previous results, we have that

- if $G^k \geq y_i/2$, then h_i^k has a *unique* fixed point, $\bar{\alpha}_i = 0$, which is attracting,[3]
- if $G^k < y_i/2$, then h_i^k has two fixed points: 0 which is repelling and $\bar{\alpha}_i > 0$ (in (18)), which is attracting.

Proof of Proposition 3

Let $\tau_i = |\lambda\theta_i^k - y_i + \frac{1}{2}| \leq 1/2$ (the upper bound comes from $\operatorname{dom}(f_i^*(-\lambda\cdot)) = [(y_i-1)/\lambda, y_i/\lambda]$). Then we get from [12, Proposition 23, see proof] (with $r = \sqrt{2G^k}/\alpha$) that

$$h_i^k(\alpha) = \begin{cases} 4\lambda^2 & \text{if } \alpha \leq \frac{2\lambda^2 G^k}{\tau_i^2} \\ \frac{4\lambda^2}{1-4\left(\tau_i - \lambda\sqrt{\frac{2G^k}{\alpha}}\right)^2} & \text{if } \alpha \geq \frac{2\lambda^2 G^k}{\tau_i^2} \end{cases}$$

$$(h_i^k)'(\alpha) = \begin{cases} 0 & \text{if } \alpha \leq \frac{2\lambda^2 G^k}{\tau_i^2} \\ \frac{16\lambda^3\sqrt{2G^k}\left(\tau_i - \lambda\sqrt{\frac{2G^k}{\alpha}}\right)}{\alpha^{\frac{3}{2}}\left(1-4\left(\tau_i - \lambda\sqrt{\frac{2G^k}{\alpha}}\right)^2\right)^2} & \text{if } \alpha \geq \frac{2\lambda^2 G^k}{\tau_i^2} \end{cases}$$

Note that $(h_i^k)'$ is continuous at $\frac{2\lambda^2 G^k}{\tau_i^2}$ and thus $h_i^k \in C^1$.

Then, from the definition of h_i^k, we distinguish two cases.

• $\bar{\alpha}_i = 4\lambda^2$ is a fixed point of h_i^k if

$$4\lambda^2 \leq 2\lambda^2 G^k/\tau_i^2 \iff G^k \geq 2\tau_i^2.$$

Moreover it is attracting (Lemma 1 with $(h_i^k)'(\bar{\alpha}_i) = 0 < 1$).

• Other fixed points of h_i^k are solutions of

$$\alpha(1-4\tau_i^2) + \sqrt{\alpha}8\tau_i\lambda\sqrt{2G^k} - 4\lambda^2(2G^k+1) = 0. \tag{19}$$

– For $\tau_i = \frac{1}{2}$, (19) is a linear equation with solution

$$\sqrt{\bar{\alpha}_i} = \lambda(2G^k+1)/\sqrt{2G^k}. \tag{20}$$

Then, one can check that

$$\bar{\alpha}_i > 8\lambda^2 G^k \iff G^k < 2\tau_i^2 = \frac{1}{2}$$

and $(h_i^k)'(\bar{\alpha}_i) = \frac{1}{2} - G^k < 1$ (i.e., $\bar{\alpha}_i$ is attracting).

– For $\tau_i < \frac{1}{2}$, (19) is a quadratic equation in $\sqrt{\alpha}$ with solutions

$$\sqrt{\alpha} = \frac{-4\tau_i\lambda\sqrt{2G^k} \pm 2\lambda\sqrt{2G^k + 1 - 4\tau_i^2}}{1-4\tau_i^2}. \tag{21}$$

As $\tau_i < \frac{1}{2}$, we have $1-4\tau_i^2 > 0$ and $4\tau_i\lambda\sqrt{2G^k} < 2\lambda\sqrt{2G^k} < 2\lambda\sqrt{2G^k + 1 - 4\tau_i^2}$, showing that the solution with the plus sign is admissible. Denoting $\bar{\alpha}_i$ this solution, we have

$$\bar{\alpha}_i > 2\lambda^2 G^k/\tau_i^2 \iff G^k < 2\tau_i^2.$$

Moreover, one can show that (see details in [14, Sec. 1.4])

$$(h_i^k)'(\bar{\alpha}_i) = \frac{\sqrt{2G^k}(2\tau_i\sqrt{2G^k + 1 - 4\tau_i^2} - \sqrt{2G^k})}{1-4\tau_i^2}. \tag{22}$$

is bounded by $\frac{1}{2}(1 - \sqrt{1-4\tau_i^2})$ when $G^k < 2\tau_i^2$. Then, with $\tau_i < \frac{1}{2}$, we get that $(h_i^k)'(\bar{\alpha}_i) < \frac{1}{2}$. This shows (with Lemma 1) that $\bar{\alpha}_i$ is attracting. Combining all these disjoint cases complete the proof.

[3] When $G^k = y_i/2$, while it is clear from the proof that 0 is the unique fixed-point, the fact that it is attracting does not come from Lemma 1 (as here $(h_i^k)'(0) = 1$). It comes from the fact that here $h_i^k(\alpha) < \alpha\ \forall \alpha \neq 0$.