1 Supplementary Material

This document contains supplementary technical content and experimental results for the following paper:

C. F. Dantas, E. Soubies, and C. Févotte, "Sphere Refinement in Gap Safe Screening," *IEEE Signal Processing Letters*, 2023.

1.1 Useful quantities and definitions

(Local) strong concavity

Definition 1. *Let $D_\lambda(\boldsymbol{\theta}) := -\sum_{i=1}^m f_i^*(-\lambda\theta_i)$ be twice differentiable, then it is $\alpha_{\mathcal{S}}$-strongly concave on $\mathcal{S} \subset \mathbb{R}^m$ if*

$$0 < \alpha_{\mathcal{S}} \leq \min_{i\in[m]} \inf_{\boldsymbol{\theta}\in\mathcal{S}\cap\operatorname{dom} f_i^*(-\lambda\cdot)} \lambda^2 f_i^{*\prime\prime}(-\lambda\theta_i), \tag{1}$$

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

Projection over the best safe region (line 7 in Alg. 2)

For completeness, we provide the formula for the projection over the current best safe region $\mathcal{S}_\mathrm{b}$ performed at line 7 in Algorithm 2.
For $\mathcal{S}_\mathrm{b} = \mathcal{B}(\boldsymbol{\theta}_{\mathcal{S}_\mathrm{b}}, r_{\mathcal{S}_\mathrm{b}})$ an ℓ_2-ball we have:

$$P_{\mathcal{S}_\mathrm{b}}(\tilde{\boldsymbol{\theta}}) = \begin{cases} \tilde{\boldsymbol{\theta}} & \text{if } \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathcal{S}_\mathrm{b}}\|_2 \leq r_{\mathcal{S}_\mathrm{b}} \\ \boldsymbol{\theta}_{\mathcal{S}_\mathrm{b}} + r_{\mathcal{S}_\mathrm{b}} \frac{(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathcal{S}_\mathrm{b}})}{\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathcal{S}_\mathrm{b}}\|_2} & \text{otherwise.} \end{cases}$$

Note that when $\tilde{\boldsymbol{\theta}} \in \mathcal{S}_\mathrm{b}$, no action is required (first case above).
Finally, when $\mathcal{S}_\mathrm{b} = \Delta_{\mathbf{A}}$ (at initial iterations), then the projection step is also unnecessary since $\tilde{\boldsymbol{\theta}} \in \Delta_{\mathbf{A}}$ (line 6 in Algorithm 2) i.e., $\tilde{\boldsymbol{\theta}}$ is dual feasible. In that case, we simply have $P_{\mathcal{S}_\mathrm{b}}(\tilde{\boldsymbol{\theta}}) = \tilde{\boldsymbol{\theta}}$.

1.2 Extended proof of Proposition 1

Definition 2 (Attracting and repelling fixed points)**.** *Let $\bar{\mathbf{x}} = f(\bar{\mathbf{x}})$ be a fixed point of $f : \mathbb{R}^N \to \mathbb{R}$ and define the fixed-point iteration as $f^n = f \circ f \circ \cdots \circ f$ (n times). Then, $\bar{\mathbf{x}}$ is said to be*

1. attracting *if there is an open neighborhood $\mathcal{X} \ni \bar{\mathbf{x}}$ such that $\forall \mathbf{x}_0 \in \mathcal{X}$, $f^n(\mathbf{x}_0) \underset{n\to\infty}{\to} \bar{\mathbf{x}}$*
2. repelling *if there is an open neighborhood $\mathcal{X} \ni \bar{\mathbf{x}}$ such that, $\forall \mathbf{x}_0 \in \mathcal{X}$, $\exists n > 0$, $f^n(\mathbf{x}_0) \notin \mathcal{X}$.*

These definitions can be found in references [15] and [16] of the main paper.

Below, we provide details on the three points of the proof of Proposition 1.

1. *Link with fixed point iteration.* Given the definition of the strong concavity bound in eq. (1), the loop over j at lines 15-18 of Alg. 1 can be rewritten more explicitly as follows:

$$\begin{aligned} r_j &\leftarrow \sqrt{2\,\mathrm{Gap}_\lambda(\mathbf{x}^k, \boldsymbol{\theta}^k)/\alpha_{j-1}} \\ \alpha_j &\leftarrow \min_{i\in[m]} \inf_{\theta_i \in \mathcal{B}(\theta_i^k, r_j)\cap\operatorname{dom} f_i^*(-\lambda\cdot)} \lambda^2 f_i^{*\prime\prime}(-\lambda\theta_i)\,. \end{aligned}$$

By plugging the expression of r_j in the update of α, we obtain

$$\alpha_j \leftarrow \min_{i\in[m]} \inf_{\substack{|\theta_i-\theta_i^k|\leq\sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}^k,\boldsymbol{\theta}^k)}{\alpha_{j-1}}} \\ \theta_i\in\mathrm{dom}\, f_i^*(-\lambda\cdot)}} \lambda^2 f_i^{*\prime\prime}(-\lambda\theta_i) \quad \Big(:= h^k(\alpha_{j-1})\Big)$$

The right-hand side of the above expression corresponds precisely to our definition of the function h^k in Proposition 1 in the main paper. This shows the equivalence between the loop in lines 15-18 and the fixed point iteration $\alpha_j = h^k(\alpha_{j-1})$.

2. *Convergence.* We know from [12, Proposition 7] that the refinement loop at lines 15-18 of Alg.1 builds a sequence of nested Gap Safe spheres (i.e., with decreasing radius), all centered in $\boldsymbol{\theta}^k$. With the previous point, this means that the fixed point iteration $\alpha_j = h^k(\alpha_{j-1})$ converges and that the convergence point is a fixed point of h^k. More precisely, it satisfies $\bar{\alpha}^k = \alpha_{\mathcal{B}}$ with $\mathcal{B} = \mathcal{B}(\boldsymbol{\theta}^k, \sqrt{\frac{2G^k}{\bar{\alpha}^k}})$.

3. *Non repelling fact.* Using again [12, Proposition 7] we get that the generated sequence $(\alpha_j)_j$ satisfies $\alpha_0 \leq \alpha_1 \leq \cdots \leq \bar{\alpha}^k$. As such, defining $\mathcal{X} = (0, \bar{\alpha}^k)$, we have shown that there is at least one point in $\mathcal{X}$ (i.e., α_0) from which all the iterates generated by the fixed point iteration $\alpha_j = h^k(\alpha_{j-1})$ belongs to $\mathcal{X}$. From Definition 2, this shows that $\bar{\alpha}^k$ is non repelling.

1.3 Properties and visualization of the fixed-point equation

In this section, we complete Proposition 1 of the main paper with some general properties of the function h^k, as well as some illustrative graphs.

Proposition 1. *Some properties of the function h^k can be inferred from its definition:*

1. *It takes only non-negative values.*
2. *It is non-decreasing.*
3. *It is continuous.*
4. *It has a horizontal asymptote, i.e. $h^k(\alpha) \to C < \infty$ as $\alpha \to \infty$.*
5. *$h^k(0) = \alpha_{\mathbb{R}^m}$, i.e. the value of $h^k(\alpha)$ at $\alpha = 0$ corresponds to the global strong-concavity constant of the the dual function D_λ. In particular, we have $h^k(0) = 0$ when D_λ is not globally strongly concave.*

Proof. We prove each statement independently.

1. Because the f_i^* are convex functions, the $f_i^{*\prime\prime}$ take only non-negative values and, as a direct consequence, so does h^k.
2. h^k is non-decreasing, since it is the inf of a family of function $(f_i^{*\prime\prime})_i$ on a ball with radius $\sqrt{2G/\alpha}$. As α increases, the radius decreases and the inf can only increase.
3. First, note that the $f_i^{*\prime\prime}$ are continuous by assumption (D_λ twice differentiable). Then, the continuity of h^k stems from the fact that it is the inf of a family of continuous function $(f_i^{*\prime\prime})_i$ on a ball with radius varying continuously on α.

4. This entails from the definition of h^k, since h^k reduces to $\min_{i\in[m]} \lambda^2 f_i^{*\prime\prime}(-\lambda\theta_i)$ when $\alpha \to \infty$ (i.e. when the radius tends to zero).

5. As $\alpha \to 0$, the radius $r = \sqrt{2G/\alpha}$ of the ball in which the strong concavity constant is calculated tends to infinity.

□

Proposition 2. *Under the working assumptions of the paper, h^k always admits at least one fixed point.*

Proof. First of all, if $h^k(0) = 0$ then 0 is a fixed point of h^k. Now let $h^k(0) > 0$ and assume that h^k does not admit a fixed point. From Proposition 1, this means that h^k is a non-decreasing, continuous, and non-negative function that is always above the identity line (i.e., $h^k(\alpha) > \alpha$). This contradicts the fact that h^k has an horizontal assymptote. Hence h^k always admits at least one fixed point. □

Illustrations for the Kullback-Leibler case: The function h^k associated to the ℓ_1 regularized Kullback-Leibler regression is displayed in Figure 1 along with its derivatives. We distinguish two relevant cases: 1) Gap $< y/2$ (top) and 2) Gap $\geq y/2$ (bottom). In both cases we have that $h(0) = 0$, which is expected since $h(0)$ corresponds to the global strong concavity bound (zero in the KL case). Note that in the latter case (bottom graph) the derivative at the origin is smaller than 1 ($h'(0) < 1$) and keeps decreasing since the function is fully concave. This implies that, in this case, the curves remain below the identity line and $h(0) = 0$ is the unique fixed-point. In the former case (top graph), the derivative at the origin is greater than 1 ($h'(0) > 1$) but then keeps decreasing (as $h''(\alpha) < 0, \forall \alpha > 0$). It becomes quite clear that, in this case, $h(\alpha)$ will eventually cross the identity line (shown in green) to produce an attractive fixed point $\bar{\alpha}$ with $h'(\bar{\alpha}) < 1$.

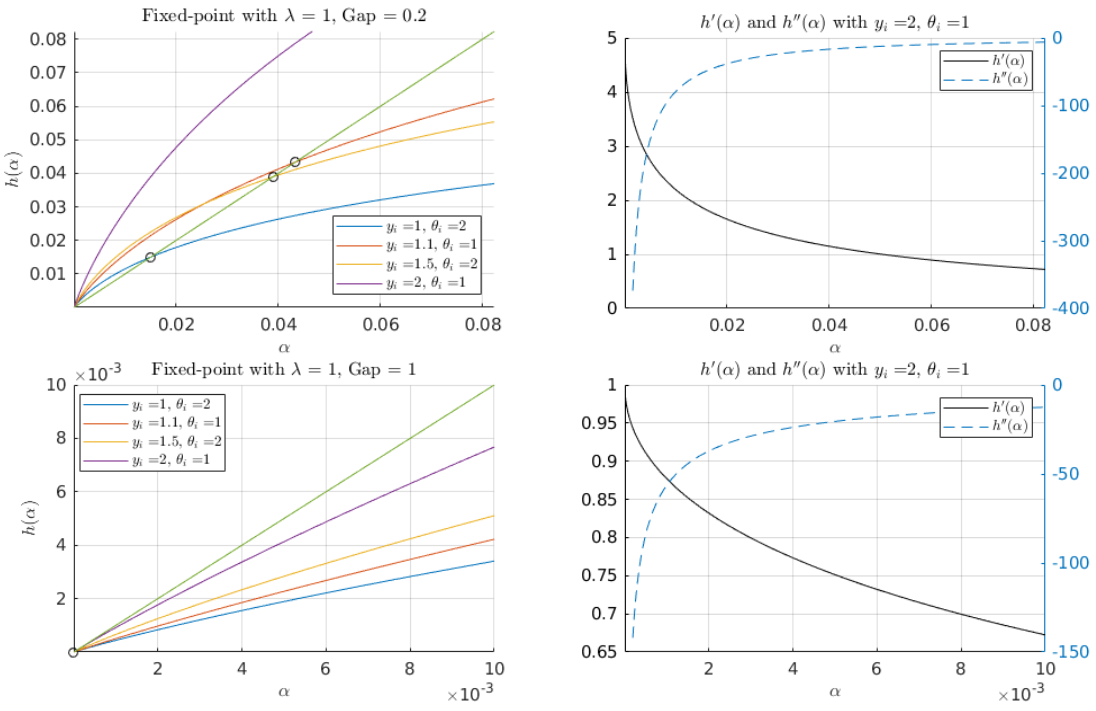

Figure 1: Visualizing the fixed-point equation for the ℓ_1-regularized logistic regression case.

Illustrations for the Logistic case: The function h^k associated to the ℓ_1 regularized logistic regression is displayed in Figure 2. Note that the function starts with a constant portion with value $4\lambda^2$, which corresponds to the global strong concavity bound $\alpha_{\mathbb{R}^m}$ for the logistic function. It then (potentially) has a brief convex portion shown in red, followed by a concave region. The identity line is shown in dark red. The first derivative remain below 1 at all times, which is in line with our analytical results proving the fixed-points to be attractive. We show the behavior of the function h for different values of $t := |\lambda\theta_i^k - y_i + \frac{1}{2}| \leq 1/2$. As predicted by eq. (14) in the paper, the fixed point equals $4\lambda^2$ (constant part of the function h) when $t \leq \sqrt{\text{Gap}/2}$ (≈ 0.22 in this example). For higher values of t (with the theoretical limit of $t \leq 1/2$) we have fixed-points necessarily with higher values than $4\lambda^2$.

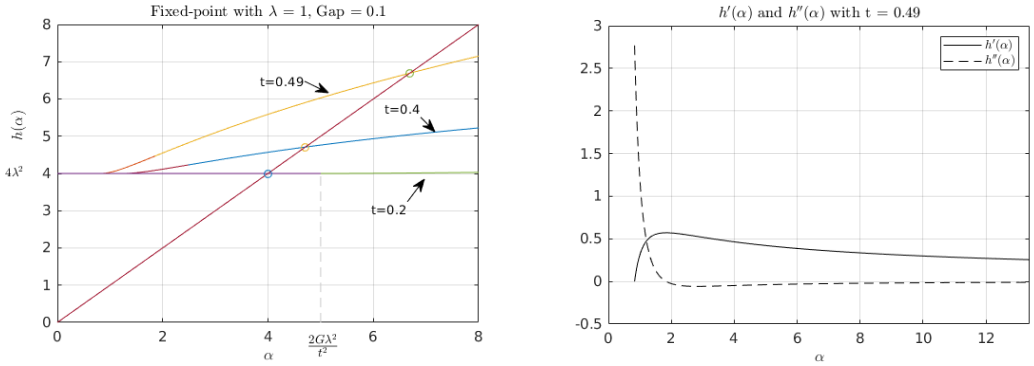

Figure 2: Visualizing the fixed-point equation for the ℓ_1-regularized logistic regression case.

1.4 Omitted details on the proof of Proposition 2

Derivation of the expression of $(h_i^k)'(\bar{\alpha}_i)$ in the case where $G^k < 2\tau_i^2$ and $\tau_i < \frac{1}{2}$: In the case where $G^k < 2\tau_i^2$ and $\tau_i < \frac{1}{2}$, we want to show that

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

where $\bar{\alpha}_i$ is given by the equation:

$$\sqrt{\bar{\alpha}_i} = \frac{-4\tau_i\lambda\sqrt{2G^k} \pm 2\lambda\sqrt{2G^k + 1 - 4\tau_i^2}}{1 - 4\tau_i^2}. \tag{3}$$

Proof. From (3) (dropping the dependencies in i and k), we get

$$\begin{aligned}
\left(\tau - \lambda\sqrt{\frac{2G}{\alpha}}\right) &= \frac{\tau\left(2\chi\sqrt{2G + 1 - 4\tau^2} - 4\tau\chi\sqrt{2G}\right) - (1 - 4\tau^2)\chi\sqrt{2G}}{2\chi\sqrt{2G + 1 - 4\tau^2} - 4\tau\chi\sqrt{2G}} \\
&= \frac{2\tau\sqrt{2G + 1 - 4\tau^2} \cancel{- 4\tau^2\sqrt{2G}} - \sqrt{2G} \cancel{+ 4\tau^2\sqrt{2G}}}{2\sqrt{2G + 1 - 4\tau^2} - 4\tau\sqrt{2G}} \\
&= \frac{1}{2}\left(\frac{2\tau\sqrt{2G + 1 - 4\tau^2} - \sqrt{2G}}{\sqrt{2G + 1 - 4\tau^2} - 2\tau\sqrt{2G}}\right)
\end{aligned}$$

We can then derive

$$1-4\left(\tau-\lambda\sqrt{\frac{2G}{\alpha}}\right)^2=\frac{\left(\sqrt{2G+1-4\tau^2}-2\tau\sqrt{2G}\right)^2-\left(2\tau\sqrt{2G+1-4\tau^2}-\sqrt{2G}\right)^2}{\left(\sqrt{2G+1-4\tau^2}-2\tau\sqrt{2G}\right)^2}$$

where one can see that the mixed products in the numerator will simplify, leading to

$$\begin{aligned}1-4\left(\tau-\lambda\sqrt{\frac{2G}{\alpha}}\right)^2&=\frac{\cancel{2G}+1-4\tau^2\cancel{+8\tau^2G}-4\tau^2(\cancel{2G}+1-4\tau^2)\cancel{-2G}}{\left(\sqrt{2G+1-4\tau^2}-2\tau\sqrt{2G}\right)^2}\\&=\frac{1-4\tau^2-4\tau^2(1-4\tau^2)}{\left(\sqrt{2G+1-4\tau^2}-2\tau\sqrt{2G}\right)^2}\\&=\frac{(1-4\tau^2)^2}{\left(\sqrt{2G+1-4\tau^2}-2\tau\sqrt{2G}\right)^2}\end{aligned}$$

We also get from (3) that

$$\alpha^{\frac{3}{2}}=(\sqrt{\alpha})^3=\frac{(2\lambda\sqrt{2G+1-4\tau^2}-4\tau\lambda\sqrt{2G})^3}{(1-4\tau^2)^3}=8\lambda^3\left(\frac{(\sqrt{2G+1-4\tau^2}-2\tau\sqrt{2G})^3}{(1-4\tau^2)^3}\right).$$

Combining these equation we can obtain both the numerator and the denominator of $h'(\alpha)$.

$$\begin{aligned}\text{Num}&=8\lambda^3\sqrt{2G}\left(\frac{2\tau\sqrt{2G+1-4\tau^2}-\sqrt{2G}}{\sqrt{2G+1-4\tau^2}-2\tau\sqrt{2G}}\right)\\\text{Den}&=8\lambda^3\left(\frac{(\sqrt{2G+1-4\tau^2}-2\tau\sqrt{2G})^3}{(1-4\tau^2)^3}\right)\frac{(1-4\tau^2)^4}{\left(\sqrt{2G+1-4\tau^2}-2\tau\sqrt{2G}\right)^4}\\&=\frac{8\lambda^3(1-4\tau^2)}{\left(\sqrt{2G+1-4\tau^2}-2\tau\sqrt{2G}\right)}\end{aligned}$$

Finally,

$$h'(\alpha)=\frac{\text{Num}}{\text{Den}}=\frac{\sqrt{2G}(2\tau\sqrt{2G+1-4\tau^2}-\sqrt{2G})}{1-4\tau^2}\tag{4}$$

which concludes the proof □

Proof that $(h_i^k)'(\bar{\alpha}_i)$ is bounded by $\frac{1}{2}(1-\sqrt{1-4\tau_i^2})<\frac{1}{2}$ over $[0,2\tau_i^2]$:

Proof. To do so, let us study $(h_i^k)'$ in (4) as a function of G. We define,

$$f(G)=\frac{\sqrt{2G}(2\tau\sqrt{2G+1-4\tau^2}-\sqrt{2G})}{1-4\tau^2}$$

from which we get

$$f'(G)=\frac{1}{1-4\tau^2}\left(\frac{8\tau G+2\tau(1-4\tau^2)-2\sqrt{2G}\sqrt{2G+1-4\tau^2}}{\sqrt{2G}\sqrt{2G+1-4\tau^2}}\right)$$

The sign if this derivative is governed by the numerator. Then,

$$\begin{aligned}
&4\tau G + \tau(1-4\tau^2) - \sqrt{2G}\sqrt{2G+1-4\tau^2} = 0 \\
\iff &(4\tau G + \tau(1-4\tau^2))^2 = 2G(2G+1-4\tau^2) \\
\iff &16\tau^2 G^2 + 8\tau^2(1-4\tau^2)G + \tau^2(1-4\tau^2)^2 = 4G^2 + 2(1-4\tau^2)G \\
\iff &4(1-4\tau^2)G^2 + 2(1-4\tau^2)^2 G - \tau^2(1-4\tau^2)^2 = 0
\end{aligned}$$

This is a quadratic equation whose positive solution is given by

$$G_{\max} = \frac{\sqrt{1-4\tau^2} - (1-4\tau^2)}{4}$$

Hence, we get that f is increasing over $[0, G_{\max}]$ and decreasing over $[G_{\max}, 2\tau^2]$. In order to find an upper-bound of $h'(\alpha)$ above, we thus have to bound $f(G_{\max})$:

$$\begin{aligned}
f(G_{\max}) &= \frac{\frac{2\tau}{\sqrt{2}}\sqrt{\sqrt{1-4\tau^2} - (1-4\tau^2)}\frac{1}{\sqrt{2}}\sqrt{\sqrt{1-4\tau^2} - (1-4\tau^2) + 2(1-4\tau^2)}}{1-4\tau^2} - \frac{\sqrt{1-4\tau^2}}{2(1-4\tau^2)} + \frac{1}{2} \\
&= \frac{\tau\sqrt{\sqrt{1-4\tau^2} - (1-4\tau^2)}\sqrt{\sqrt{1-4\tau^2} + (1-4\tau^2)}}{1-4\tau^2} - \frac{\sqrt{1-4\tau^2}}{2(1-4\tau^2)} + \frac{1}{2} \\
&= \frac{\tau\sqrt{(1-4\tau^2) - (1-4\tau^2)^2}}{1-4\tau^2} - \frac{\sqrt{1-4\tau^2}}{2(1-4\tau^2)} + \frac{1}{2} \\
&= \frac{\tau\sqrt{4\tau^2(1-4\tau^2)}}{1-4\tau^2} - \frac{\sqrt{1-4\tau^2}}{2(1-4\tau^2)} + \frac{1}{2} \\
&= \frac{\sqrt{1-4\tau^2}}{(1-4\tau^2)}\left(2\tau^2 - \frac{1}{2}\right) + \frac{1}{2} \\
&= \frac{1}{2}\left(1 - \sqrt{1-4\tau^2}\right) < \frac{1}{2}
\end{aligned}$$

where we used the fact that $\tau < \frac{1}{2}$ to obtain the last inequality. □

1.5 Execution times comparison

Execution time results reported in Table 1 show accelerations of about 5, 6 and 13 times for coordinate descent, proximal gradient and majorize-minimization solvers respectively. Also note that Alg. 2 allows significantly reduce the time consecrated to the screening tests when compared to the iterative approach in Alg. 1. However, these screening times remain quite small compared to the overall execution times.

1.6 Number of refinement iterations

In Figure 3 we show the number of refinement iterations per solver iteration for the same scenarios shown in Table 1. Results for the logistic regression case are given in Figure 4. One can see that the number of refinement iterations revolves around 5 or 10 (respectively for $\varepsilon_r = 10^{-3}$ or 10^{-5}) before it stabilizes to 2 or 1. Nevertheless, the total accumulated number of refinement iterations remain considerable in most cases.

Effect of poor initialization

A higher number of refinement iterations tends to occur when, for some reason, there is a big margin of improvement for the current value of α. To simulate one such scenario, we purposely

Table 1: Execution times for KL-L1 regression on TasteProfile data, with $\lambda/\lambda_{\max} = 10^{-3}$. Here, τ denotes the screening frequency (e.g. $\tau = 10$ means that screening tests are performed every 10 solver iterations). The number of refinement iterations avoided by Alg. 2 is also reported.

			Total time [s]			Screening time [s]		Refinement it.
	ε_r	τ	No screen.	Alg. 1	Alg. 2	Alg. 1	Alg. 2	Total
Coord Desc.	10^{-3}	1	27.23	5.93	5.92	1.64e-2	1.08e-2	113
		10	29.23	4.06	4.03	9.27e-3	4.75e-3	17
	10^{-5}	1	27.30	6.04	5.95	1.97e-2	1.03e-2	192
		10	27.09	4.15	4.11	1.06e-2	4.64e-3	27
Prox. Grad.	10^{-3}	1	15.97	2.23	2.22	3.82e-2	3.11e-2	475
		10	15.70	2.50	2.44	6.41e-3	4.54e-3	66
	10^{-5}	1	16.24	2.26	2.20	3.90e-2	3.14e-2	576
		10	10.89	2.31	2.30	6.99e-3	4.93e-3	89
Maj. Min.	10^{-3}	1	13.97	1.11	1.00	8.97e-2	6.08e-2	314
		10	12.92	0.74	0.66	1.85e-2	1.08e-2	135
	10^{-5}	1	13.11	1.11	1.02	1.18e-1	6.20e-2	3593
		10	13.08	0.78	0.69	2.15e-2	1.19e-2	578

initialize α with a smaller value in Figure 5 (in the right graph, α initialization is divided by a factor of 100). One can verify that the number of refinement iterations indeed goes up considerably (about an order of magnitude). In such cases, the proposed analytic approach becomes particularly interesting.

Remark 1. *It is important to emphasize that, contrarily to the iterative variant, the analytic approach does not require the initialization of the strong concavity bound α. This can be decisive in some cases where a global bound (or a bound over the feasible set) is not readily available. Inversely, when a closed-form solution of the refinement fixed-point equation is not available, the iterative variant can be used. When neither is available (global initialization or closed-form solution) then the iterative variant can be deployed with a very low initialization for α, but then the number of required refinement iterations will typically be very high.*

Warm start

When a full regularization path is being solved sequentially, starting with a high regularization and gradually reducing its value, the next problem instance can be initialized with the previous solution for a warm start. In such cases, the primal and dual estimates can be already quite accurate in initial iterations, while α has to be reset to its pessimistic global bound. One would expect a large number of refinement steps to be necessary in the initial solver iterations until α estimates catch-up. This hypothesis was indeed verified empirically, as reported in Table 2. Note that the number of refinement iterations increases with a finer grid and so does the time saved by the analytic approach. Indeed, a finer grid implies that the warm start initialization is more accurate and, therefore, the α initialization is comparatively worse.

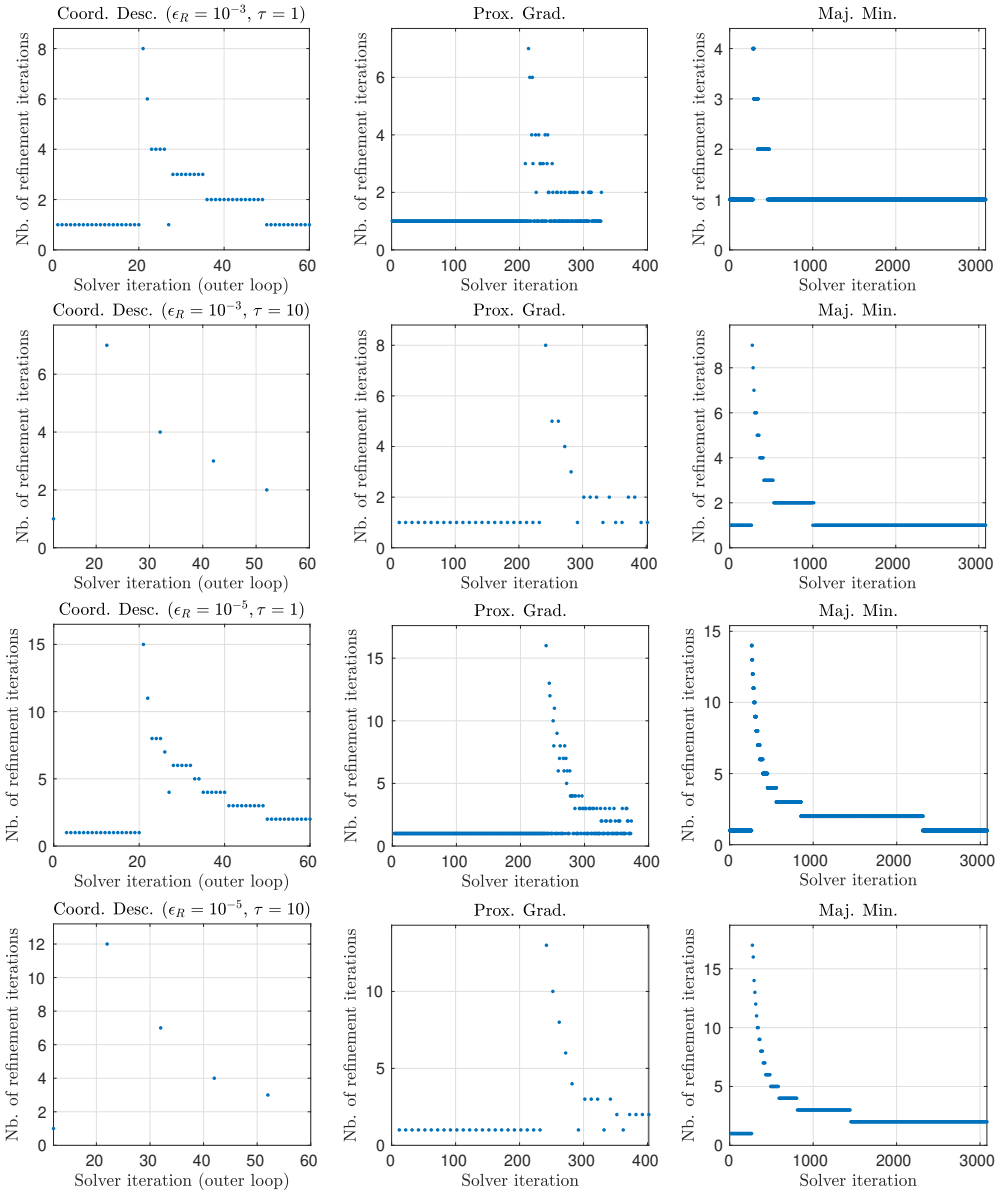

Figure 3: Number of refinement iterations in Alg. 1 per solver iteration in the KL-L1 regression problem with $\lambda/\lambda_{\max} = 10^{-3}$ and $\varepsilon_r \in \{10^{-3}, 10^{-5}\}$, $\tau \in \{1, 10\}$ (same as in Table 1).

Table 2: Screening times and refinement iterations for a KL-L1 regularization path with warm start and $\lambda/\lambda_{\max} \in [10^{-3}, 1)$ at different grid resolutions (i.e. the number of regularization values taken logarithmically-spaced in the grid) with $\varepsilon_r = 10^3$ and $\tau = 1$.

		Screening time [s]			Refinement it.	
	Grid resolution	Alg. 1	Alg. 2	Ratio (2/1)	Total	Per λ
Coord. Desc.	20	2.5e-2	1.5e-2	0.60	237	11.9
	100	13.3e-2	6.9e-2	0.52	1521	15.2
Prox. Grad.	20	1.9e-2	1.4e-2	0.74	319	15.9
	100	7.9e-2	5.4e-2	0.68	1792	17.9

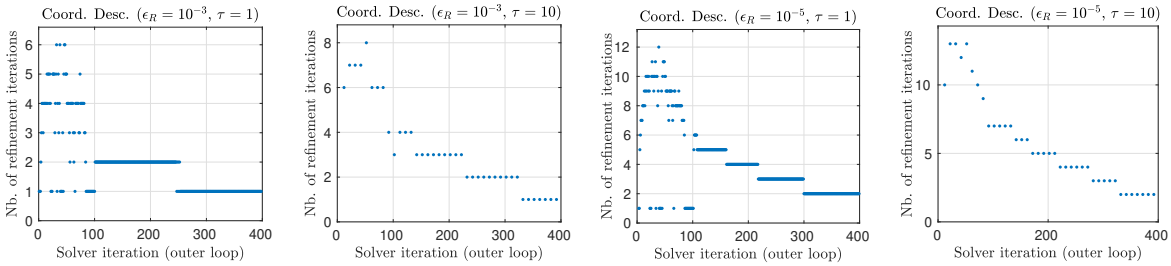

Figure 4: Number of refinement iterations in Alg. 1 per solver iteration in the Logistic-L1 regression problem with $\lambda/\lambda_{\max} = 10^{-3}$ and $\varepsilon_r \in \{10^{-3}, 10^{-5}\}$, $\tau \in \{1, 10\}$.

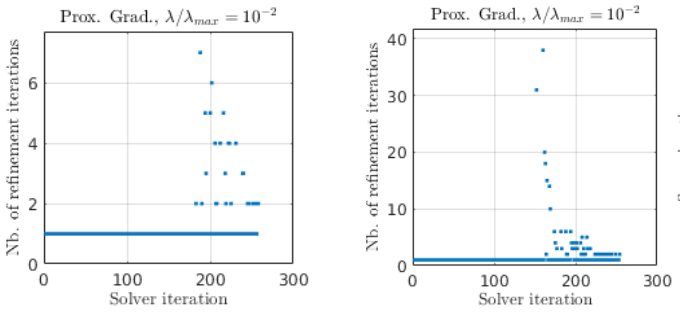

Figure 5: Effect of a poor initialization of α in the number of refinement iterations. Left: standard initialization. Right: degraded initialization (by a factor of 100).

1.7 Cases (improvement, indecisive, no improvement) distribution

Sensitivity to regularization

In Tables 3 and 4 we show the proportion of occurrence of each of three possible cases (improvement, indecisive and no improvement) at three different regularization regimes $\lambda/\lambda_{\max} \in \{10^{-1}, 10^{-2}, 10^{-3}\}$.

Table 3: Cases distribution (Improvement; No improvement; Indecisive) for KL-L1 regression on NIPSpapers data, with regularization $\lambda/\lambda_{\max} \in \{10^{-1}, 10^{-2}, 10^{-3}\}$.

		$\lambda/\lambda_{\max} = 10^{-1}$		$\lambda/\lambda_{\max} = 10^{-2}$		$\lambda/\lambda_{\max} = 10^{-3}$	
		$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$
Coord. Desc.	Improv.	0	44.5	0	49.6	0	48.6
	No-Improv.	52.2	0	49.6	0	51.4	0
	Indec.	2.2	1.1	0.8	0	0	0
Prox. Grad.	Improv.	0	98.5	0	100	0	92.5
	No-Improv.	1.5	0	0	0	0	0
	Indec.	0	0	0	0	7.5	0

Because the results are quite robust to the regularization parameter, from this point on (and in the paper) we report average results on 100 different regularizations on a grid $\lambda/\lambda_{\max} \in [10^{-3}, 1)$.

Table 4: Cases distribution for Logistic-L1 regression on Leukemia data.

		$\lambda/\lambda_{\max} = 10^{-1}$		$\lambda/\lambda_{\max} = 10^{-2}$		$\lambda/\lambda_{\max} = 10^{-3}$	
		$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$
Coord. Desc.	Improv.	0	86.8	0	89.8	0	92.9
	No-Improv.	10.6	0	8.2	0	5.3	0
	Indec.	1.3	1.3	1.4	0.7	1.3	0.5

Sensitivity to data matrix

Results with TasteProfile and 20newsgrous datasets are shown in Table 5 (to be compared with NIPSpapers results in Table 1 in the paper).

Table 5: KL-L1 regression on different datasets.

	Proximal Gradient				Coordinate Descent			
	Taste Profile		20 News Groups		Taste Profile		20 News Groups	
	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$
Improv.	0	38.7 (5.6)	0	47.5 (7.1)	0	98.4 (3.7)	0	99.2 (2.5)
No-Imp.	60.2 (5.6)	0	48.8 (5.9)	0	1.4 (3.5)	0	0.3 (1.2)	0
Indec.	0.5 (0.9)	0.6 (1.1)	1.8 (2.3)	1.9 (2.2)	0.1 (0.4)	0.1 (0.6)	0.3 (0.8)	0.2 (0.8)

Sensitivity to input vector

In Table 6 we show the results for different input vectors $\mathbf{y}$, which, in the this archetypal analysis setup, is a randomly select sample extracted from the data matrix. All realizations shown below are different from the ones used in Table 1 in the paper.[1]

Table 6: KL-L1 regression with different input vectors.

	Proximal Gradient				Coordinate Descent			
	Realization 1		Realization 2		Realization 1		Realization 2	
	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k \leq \alpha_{\mathcal{S}_b}$	$\bar{\alpha}^k > \alpha_{\mathcal{S}_b}$
Improv.	0	46.1 (6.7)	0	47.8 (9.8)	0	88.7 (8.5)	0	81.1 (9.3)
No-Imp.	52.5 (6.7)	0	51.1 (9.3)	0	11.1 (9.3)	0	18.6 (9.2)	0
Indec.	0.8 (1.5)	0.6 (1.0)	0.5 (1.2)	0.6 (1.2)	0	0.2 (0.9)	0.1 (0.5)	0.2 (0.6)

[1]This experiment doesn't make sense for the Logistic Regression case, in which the input vector is fixed and corresponds to the classification labels.