On L0 Bregman-Relaxations for Kullback-Leibler Sparse Regression

Mhamed Essafri, Luca Calatroni, Emmanuel Soubies

ON ℓ_0 BREGMAN-RELAXATIONS FOR KULLBACK-LEIBLER SPARSE REGRESSION

M'hamed Essafri(1), Luca Calatroni(2), Emmanuel Soubies(1)

(1) IRIT, Université de Toulouse, INP, CNRS (2) CNRS, UniCA, Inria, I3S laboratory

ABSTRACT

The resolution of optimization problems involving the ℓ_0 pseudo-norm has proven to be of importance in signal processing and machine learning applications for selecting relevant variables. Among the vast class of existing approaches dealing with the intrinsic NP-hardness of such problems, continuous (possibly non-convex) relaxations have been increasingly considered over the recent years. The notion of ℓ_0-Bregman relaxation (B-rex) has been recently introduced to construct effective relaxations of ℓ_0-regularized objectives with general data terms. These relaxations are termed *exact* in the sense that they preserve the global minimizers while removing some local minimizers. In this study, we deepen this idea further for ℓ_0-regularized Kullback-Leibler regression problems, designing a tailored B-rex. Compared to other relaxations, it further reduces the number of local minimizers of the original problem by means of a suitable analytical/geometrical modeling. To better exploit the geometry of the relaxed problem, we deploy a dedicated Bregman proximal gradient algorithm for its minimization.

Index Terms— ℓ_0-regularization, non-convex optimization, continuous relaxations, Kullback-Leibler divergence.

1. INTRODUCTION

In the context of machine learning and signal/imaging inverse problems, the Poisson observation model is often considered in scenarios involving the recording of discrete events. This is the case, for instance, of medical, biological, and astronomical imaging. Mathematically, the model reads:

$$\mathbf{y} \sim \text{Poisson}(\mathbf{A}\mathbf{x}) \tag{1}$$

where, $\mathbf{y} \in \mathbb{R}^M_{\geq 0}$ is the observation vector, $\mathbf{x} \in \mathbb{R}^N_{\geq 0}$ is the signal to retrieve, and $\mathbf{A} \in \mathbb{R}^{M\times N}_{\geq 0}$ is the (typically wide) measurement matrix ($M \ll N$). Sparse optimization approaches for (1) aim at estimating a sparse signal $\hat{\mathbf{x}}$ from $\mathbf{y}$. To enforce sparsity, the ℓ_0 pseudo-norm is here considered as regularization. As a data term, we consider the Kullback-Leibler divergence, which is known to be the natural choice from a Bayesian perspective [1]. The optimization problem of interest thus reads:

$$\hat{\mathbf{x}} \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^N_{\geq 0}} J_0(\mathbf{x}) := D_{\mathrm{KL}}(\mathbf{y}; \mathbf{A}\mathbf{x} + \mathbf{b}) + \lambda\|\mathbf{x}\|_0, \tag{2}$$

where the term $\|\mathbf{x}\|_0$ counts the number of non-zero elements of $\mathbf{x} \in \mathbb{R}^N_{\geq 0}$, and $\lambda > 0$ is a positive parameter controlling the trade-off between data fidelity and sparsity. The data term D_{KL} corresponds to the Poisson negative log-likelihood and reads:

$$D_{\mathrm{KL}}(\mathbf{y}; \mathbf{A}\mathbf{x} + \mathbf{b}) = \sum_{m=1}^{M} d_{\mathrm{KL}}(y_m; [\mathbf{A}\mathbf{x}]_m + b_m),$$

with $d_{\mathrm{KL}}(y; z) = z - y\log(z)\,,\ \forall z \in \mathbb{R}_{\geq 0}$. For $m \in [M]$ the scalars $b_m > 0$ are positive parameters which are often employed in the modeling to make the function well defined at zero. Note, however, that in several applications they represent background intensities, so that the analog to (1) becomes $\mathbf{y} \sim \text{Poisson}(\mathbf{A}\mathbf{x} + \mathbf{b})$ with $\mathbf{b} = (b_m)_{1\leq m\leq M}$, see, e.g., [2].

Minimizing the ℓ_0 pseudo-norm is the natural choice to promote sparsity of the solutions. However, it is non-continuous and non-convex, which makes Problem (2) NP-hard [3]. Many efforts have been made to solve this problem. One strategy which we will follow in this work is to relax Problem (2) by replacing the ℓ_0 term with a continuous approximation. We further require that the (generally non-convex) resulting continuous relaxation

i) preserves the global minimizers of J_0,
ii) removes many local (not global) minimizers of J_0.

Relaxations satisfying these two properties are referred to as *exact continuous relaxations*.

For a least-squares data term, an exact continuous relaxation referred to as CEL0 (continuous exact ℓ_0) has been proposed in [4]. In [5], the author demonstrated that this relaxation can be geometrically interpreted as a quadratic envelope of the ℓ_0 regularizer. Beyond the least-squares case, in [6], a class of MPEC (mathematical programs with equilibrium constraints) exact relaxations is proposed. In [7], the authors showed that the capped-ℓ_1 penalty leads to exact relaxations when the data terms is Lipschitz continuous. For weighted-ℓ_2 data terms, often used to approximate the KL divergence in applications, a weighted-CEL0 relaxation has been proposed in [8]. In [9], the authors extended the analysis carried out in [4, 5] to general (i.e., non-quadratic) data terms by using Bregman divergences. They introduced the class of ℓ_0 Bregman-relaxations (B-rex) leading to exact relaxations of (2). Note that this framework includes the CEL0 relaxation, being it associated to the case of a standard (Euclidean) Bregman geometry.

Contribution. Inspired by [9], we propose in this work an improved continuous exact relaxation for Problem (2) in the sense that it enjoys a more favorable optimization landscape with less local minimizers and wider basins of attraction (Section 3). Moreover, we propose a Bregman proximal gradient algorithm to adapt the optimization trajectory to the geometry induced by the Bregman distance considered to construct the relaxation (Section 4).

Notation. We denote by $[N] = \{1, \ldots, N\}$ the set of indices up to $N \in \mathbb{N}^*$. $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} : x \geq 0\}$ is the non-

The authors acknowledge the support of ANRs MICROBLIND (ANR-21-CE48-0008), EROSION (ANR-22-CE48-0004), TASKABILE (ANR-22-CE48-0010), and GdR ISIS project SPLIN.

negative orthant. For a vector $\mathbf{x} \in \mathbb{R}^N$, $\sigma(\mathbf{x})$ denotes its support, that is the set $\sigma(\mathbf{x}) = \{n \in [N] : x_n \neq 0\}$ and $\mathbf{x}^{(n)} = (x_1, \ldots, x_{n-1}, 0, x_{n+1}, \ldots, x_N) \in \mathbb{R}^N$. For $n \in [N]$, the vector $\mathbf{e}_n \in \mathbb{R}^N$ stands for the unit vector of the standard basis of $\mathbb{R}^N$ and $\mathbf{a}_n \in \mathbb{R}^M$ is the nth column of the matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$. We consider the simplified notations $d'_{\mathrm{KL}}(y; x + b) = (d_{\mathrm{KL}}(y; \cdot + b))'(x)$ and $d''_{\mathrm{KL}}(y; x + b) = (d_{\mathrm{KL}}(y; \cdot + b))''(x)$ to denote derivatives with respect to x.

2. BACKGROUND ON ℓ_0 BREGMAN RELAXATIONS

We recall here the main results of [9] that will be useful to the subsequent analysis. Given a family $\Psi = \{\psi_n : \mathbb{R}_{\geq 0} \to \mathbb{R}\}_{n \in [N]}$ of strictly convex, proper and twice-differentiable functions on $\mathrm{int}(\mathbb{R}_{\geq 0})$, the ℓ_0 Bregman relaxation is defined for all $\mathbf{x} \in \mathbb{R}^N_{\geq 0}$ by:

$$B_\Psi(\mathbf{x}; \lambda) = \sum_{n=1}^{N} \beta_{\psi_n}(x_n; \lambda).$$

For $x \in \mathbb{R}_{\geq 0}$, the scalar functions β_{ψ_n} read:

$$\beta_{\psi_n}(x; \lambda) = \begin{cases} \psi_n(0) - \psi_n(x) + \psi'_n(\alpha_n^+)x, & \text{if } x \in [0, \alpha_n^+] \\ \lambda, & \text{otherwise} \end{cases}, \quad (3)$$

where α_n^+ is the unique solution of $\psi_n(0) - \psi_n(x) + \psi'_n(x)x = \lambda$. With B_Ψ, a continuous relaxation of (2) can be defined by

$$J_\Psi(\mathbf{x}) = D_{\mathrm{KL}}(\mathbf{y}; \mathbf{A}\mathbf{x} + \mathbf{b}) + B_\Psi(\mathbf{x}; \lambda). \quad (4)$$

According to [9, Theorem 9], a sufficient condition for this relaxation to be exact, referred to as *concavity condition*, is given by: for all $n \in [N]$ and $\mathbf{x} \in \mathbb{R}^N_{\geq 0}$,

$$g(t) := J_\Psi(\mathbf{x}^{(n)} + t\mathbf{e}_n) \text{ is strictly concave on } (0, \alpha_n^+). \quad \text{(C1)}$$

In [9] the authors dealt with a simplified condition that decouples the concavity of $\beta_{\psi_n}(x; \lambda)$ on $(0, \alpha_n^+)$ and the convexity of D_{KL}. It is given by: for all $n \in [N]$,

$$\inf_{t \in (0, \alpha_n^+)} \psi''_n(t) > \sum_{m=1}^{M} a_{mn}^2 \sup_{z \in \mathbb{R}_{\geq 0}} d''_{\mathrm{KL}}(y_m; z + b_m). \quad \text{(C2)}$$

However, condition (C2) does not fully exploit the 1D geometrical condition described by (C1). It is therefore coarser, albeit easier to manipulate. The goal of the following section is to introduce a suitable Bregman geometry allowing us to directly leverage condition (C1), without need to lose tightness by taking infima and suprema as in (C2).

3. A TAILORED B-REX FOR SPARSE KL PROBLEMS

3.1. Proposed B-rex and Conditions for Exactness

We recall some Bregman generating functions ψ_n used in [9] to construct B_Ψ in (3). For $\gamma_n > 0$ we consider in particular $\psi^{\mathrm{L2}}_{\gamma_n}(x) = \frac{\gamma_n}{2}x^2$, the 2-power function, and, for $\epsilon > 0$ $\psi^{\mathrm{KL}}_{\gamma_n,\epsilon}(x) = \gamma_n d_{\mathrm{KL}}(1; x + \epsilon)$. For such choices, condition (C2) simplifies to

$$\gamma_n \geq \sum_{m=1}^{M} \frac{a_{mn}^2 y_m}{b_m^2}, \quad \text{(C2-}\psi^{\mathrm{L2}}_{\gamma_n}\text{)}$$

$$\frac{\gamma_n}{\epsilon^2} W^2(-\mathrm{e}^{-\kappa}) > \sum_{m=1}^{M} \frac{a_{mn}^2 y_m}{b_m^2}, \quad \text{(C2- }\psi^{\mathrm{KL}}_{\gamma_n,\epsilon}\text{)}$$

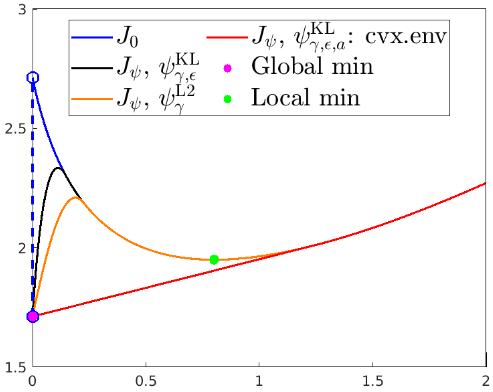


Fig. 1: Comparison between the original functional J_0 and the relaxed functional J_ψ computed for $\psi \in \left\{\psi^{\mathrm{L2}}_\gamma, \psi^{\mathrm{KL}}_{\gamma,\epsilon}, \psi^{\mathrm{KL}}_{\gamma,\epsilon,a}\right\}$, with $y = 0.7$, $a = 0.75$ and $b = 0.1$.

where $W(\cdot)$ denotes the Lambert function and $\kappa := \frac{\lambda}{\gamma_n} + 1$. To directly exploit the tighter condition expressed by (C1), we consider in this work additional parameters c_n, $n \in [N]$, in the definition of the KL generating function and define:

$$\psi^{\mathrm{KL}}_{\gamma_n,\epsilon,c_n}(x) := \gamma_n d_{\mathrm{KL}}(1; c_n x + \epsilon). \quad (5)$$

We will show that this choice allows us to better adjust the geometry of the relaxation to the one of the KL data term. The following proposition provides the conditions on the parameters to ensure the validity of (C1). The proof is provided in Appredix A.

Proposition 1. *For $n \in [N]$, let $\psi^{\mathrm{KL}}_{\gamma_n,\epsilon,c_n}$ be defined as in* (5) *with $\boldsymbol{\gamma} \in \mathbb{R}^N_{>0}$, $\epsilon > 0$, and $\mathbf{c} \in \mathbb{R}^N_{\geq 0}$ such that:*

$$c_n = \min_{m \in \sigma(\mathbf{a}_n)} a_{mn}, \quad \gamma_n > \sum_{m=1}^{M} \frac{a_{mn}^2 y_m}{c_n^2}, \quad \epsilon \leq \min_{m \in [M]} b_m, \quad (6)$$

where $\sigma(\mathbf{a}_n)$ denotes the support of the vector $\mathbf{a}_n$. Then, (C1) *holds and the functional* (4) *is an exact continuous relaxation of* (2)*.*

We highlight that conditions (6) are explicit as opposed to the condition (C2- $\psi^{\mathrm{KL}}_{\gamma_n,\epsilon}$) which requires to solve an equation involving the Lambert function. Moreover, we will describe in the following paragraph, the proposed relaxation enjoys a more favorable landscape than that obtained by $\psi^{\mathrm{L2}}_{\gamma_n}$ and $\psi^{\mathrm{KL}}_{\gamma_n,\epsilon}$.

3.2. Numerical Illustrations

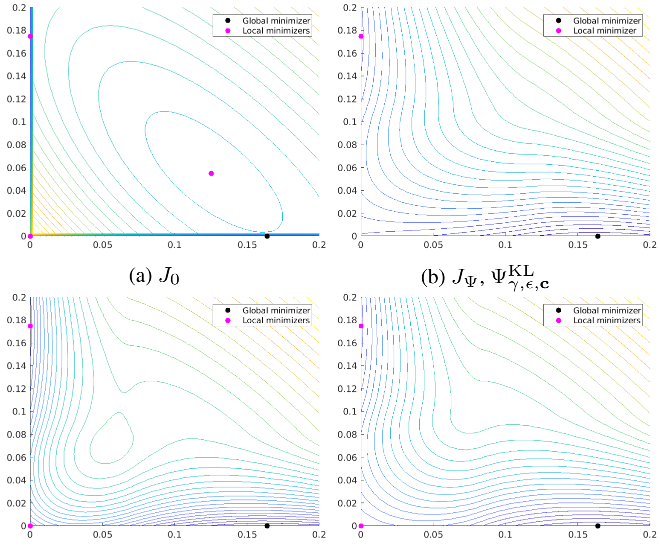
One-dimensional example. We consider the one-dimensional problem ($M = N = 1$) with $\lambda = 1$, $a > 0$ and $y \geq 0$:

$$J_0(x) = d_{\mathrm{KL}}(y, ax + b) + |x|_0.$$

We want to compare J_0 with the exact relaxations defined by

$$J_\psi(x) = d_{\mathrm{KL}}(y; ax + b) + \beta_\psi(x; 1)$$

with generating functions $\left\{\psi^{\mathrm{L2}}_\gamma, \psi^{\mathrm{KL}}_{\gamma,\epsilon}, \psi^{\mathrm{KL}}_{\gamma,\epsilon,a}\right\}$. In Figure 1, we plot the original function J_0 in blue, along with its minimizers: the global minimizer $x^* = 0$ and a local minimizer $x = \frac{y-b}{a}$. Then, we plot the three exact relaxations J_ψ obtained by choosing ψ as $\psi^{\mathrm{KL}}_{\gamma,\epsilon}$



(a) J_0 (b) J_Ψ, $\Psi^{\mathrm{KL}}_{\gamma,\epsilon,\mathbf{c}}$

(c) J_Ψ, $\Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon}$ (d) J_Ψ, $\Psi^{\mathrm{L2}}_{\boldsymbol{\gamma}}$

Fig. 2: Level lines of J_0 and J_Ψ with $\Psi \in \left\{\Psi^{\mathrm{L2}}_{\boldsymbol{\gamma}}, \Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon}, \Psi^{\mathrm{KL}}_{\gamma,\epsilon,\mathbf{c}}\right\}$ with $\mathbf{A} = [0.45, 0.8; 0.85, 0.25]$, $\mathbf{y} = [0.2; 0.22]$, $\lambda = 0.06 \times D_{\mathrm{KL}}(\mathbf{y};\mathbf{0})$ and $b = 0.1$.

(black), $\psi^{\mathrm{KL}}_{\gamma,\epsilon,a}$ (red) and $\psi^{\mathrm{L2}}_{\gamma}$ (orange). On the one hand, we observe that by considering the proposed $\psi = \psi^{\mathrm{KL}}_{\gamma,b,a}$ with parameters prescribed by condition (6), J_ψ corresponds to the convex envelope of J_0 (red). In this case, the unique global minimizer x^* of J_ψ thus coincides with the global minimizer of J_0. On the other hand, when considering J_ψ with $\psi = \psi^{\mathrm{KL}}_{\gamma,b}$ (black) or $\psi = \psi^{\mathrm{L2}}_{\gamma}$ (orange) with parameters as in conditions (C2- $\psi^{\mathrm{KL}}_{\gamma_n,\epsilon}$) and (C2-$\psi^{\mathrm{L2}}_{\gamma_n}$), respectively, a non-convex exact continuous relaxation where the global minimizer and local minimizer are not shifted is obtained.

Two-dimensional example. For a 2D illustration, we report in Figure 2 the isolevels of both J_0 and J_ψ computed for different choices of the generating functions, along with their respective minimizers. The plots show that the global minimizer of J_0 is preserved by J_Ψ for all cases. Note that the best optimization landscape is obtained by $\Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon,\mathbf{c}}$, as two local (not global) minimizers are removed in comparison with the other two relaxations which remove only one local minimizer.

Higher-dimensional examples. To assess the quality of the proposed relaxation in higher dimensions, we exploit the result [9, Theorem 3] characterizing the strict local minimizers of Problem (2) (including global ones [9, Theorem 4]) as well as [9, Proposition 10] which provides conditions for a local minimizer $\hat{\mathbf{x}}$ of J_0 to be preserved by an exact relaxation J_Ψ.

More precisely, leveraging [9, Theorem 3], we can efficiently compute all strict minimizers of J_0 by solving convex problems, since for any support $\omega \in \hat{\Omega}$ with

$$\hat{\Omega} = \bigcup_{r=0}^{M} \Omega_r, \quad \Omega_r = \{\omega \subset [N] \mid \sharp\omega = r = \mathrm{rank}(\mathbf{A}_\omega)\},$$

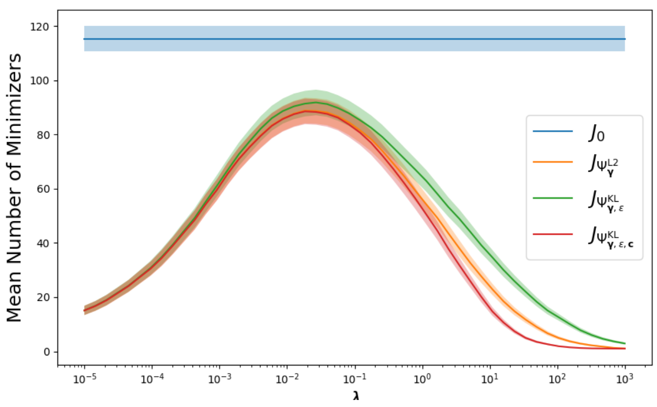


Fig. 3: Average number of strict local minimimizers of J_0 (blue) compared to their average number preserved the exact relaxations J_Ψ for $\Psi \in \{\Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon,\mathbf{c}}, \Psi^{\mathrm{L2}}_{\boldsymbol{\gamma}}, \Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon}\}$, with standard deviation (shadowed regions). Experiments are performed over 100 random generations of $\mathbf{A} \in \mathbb{R}^{5\times 10}_{\geq 0}$ and $\mathbf{y} \in \mathbb{R}^{5}_{\geq 0}$.

we get a strict local minimizer $\hat{\mathbf{x}}$ of J_0 by setting $\hat{\mathbf{x}}_{\omega^c} = \mathbf{0}$ and

$$\hat{\mathbf{x}}_\omega = \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^{\sharp\omega}_{\geq 0}} D_{\mathrm{KL}}\left(\mathbf{y}; \mathbf{A}_\omega \mathbf{z} + \mathbf{b}\right) \tag{7}$$

where $\mathbf{A}_\omega \in \mathbb{R}^{M\times\sharp\omega}_{\geq 0}$ is obtained by selecting the columns of $\mathbf{A}$ indexed by ω. Then, we can exploit [9, Proposition 10] that states that a strict local minimizer $\hat{\mathbf{x}}$ of J_0 is also a strict local minimizer of an exact relaxation J_Ψ if and only if

$$\begin{array}{l} \forall n \in \sigma(\hat{\mathbf{x}}),\ \hat{x}_n \in (\alpha_n^+, +\infty) \\ \forall n \in \sigma^c(\hat{\mathbf{x}}),\ \langle \mathbf{a}_n, (\nabla D_{\mathrm{KL}}(\mathbf{y}; \cdot + \mathbf{b}))(\mathbf{A}\mathbf{x})\rangle \in (-\infty, \ell_n^+] \end{array} \tag{8}$$

where $\ell_n^+ = \psi_n'(\alpha_n^+) - \psi_n'(0)$ and $\sigma(\hat{\mathbf{x}})$ denotes the support of $\hat{\mathbf{x}}$.

By combining these two results, we can compute the number of strict local minimizers of J_0 that are preserved by an exact relaxation J_Ψ. Given that the number of strict minimizers of J_0 grows rapidly with the dimension of the problem, we consider problems of limited size $(M, N) = (5, 10)$ and initiate these experiments by generating 100 synthetic instances of $(\mathbf{A}, \mathbf{y})$, as described in [9, section 6.3]. Then, for each instance, we compute all the strict minimizers of J_0 and identify how many are preserved by the exact relaxation computed by choosing different generating functions Ψ for 40 different values of λ in $[10^{-5}, 10^{3}]$. We report the results in Figure 3. While, for a given instance of $(\mathbf{A}, \mathbf{y})$, the number of strict local minimizers of J_0 does not change with λ (indeed (7) is independent of λ), we observe that all relaxations have less strict local minimizers than J_0 and that this number varies with λ. In particular, for very small and very large values of λ, the relaxations are equivalent in terms of eliminating strict local (not global) minimizers of J_0. However, for practical values of λ, the exact relaxation obtained with $\Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon,\mathbf{c}}$ (red) removes more strict local (not global) minimizers than those obtained with $\Psi^{\mathrm{L2}}_{\boldsymbol{\gamma}}$ (orange) and $\Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon}$ (green).

Remark 1. *Note that the number of strict minimizers of J_0 can vary from one realization of $(\mathbf{A}, \mathbf{y})$ to another due to the non-negativity constraint. Indeed, two different supports in $\hat{\Omega}$ may lead to the same strict local minimizer. This explains the non-zero standard deviation of the blue curve in Figure 3.*

To illustrate the behavior of the proposed relaxation in higher dimension, let us remark from (8) that the larger α_n^+ (resp. the smaller

ψ_n	α_n^+	ℓ_n^+
$\psi^{\mathrm{L2}}_{\gamma_n}$	$\sqrt{\frac{2\lambda}{\gamma_n}}$	$\sqrt{2\lambda\gamma_n}$
$\psi^{\mathrm{KL}}_{\gamma_n,\epsilon,c_n}$	$-\frac{1}{c_n}\left(\frac{b}{W(-\mathbf{e}^{-\kappa})}+b\right)$	$\frac{\gamma_n c_n}{b}\left(1+W(-\mathbf{e}^{\kappa})\right)$

Table 1: Quantities α_n^+ and ℓ_n^+ for different ψ_n. For the choice $\psi^{\mathrm{KL}}_{\gamma_n,\epsilon}$ we considered $c_n = 1$.

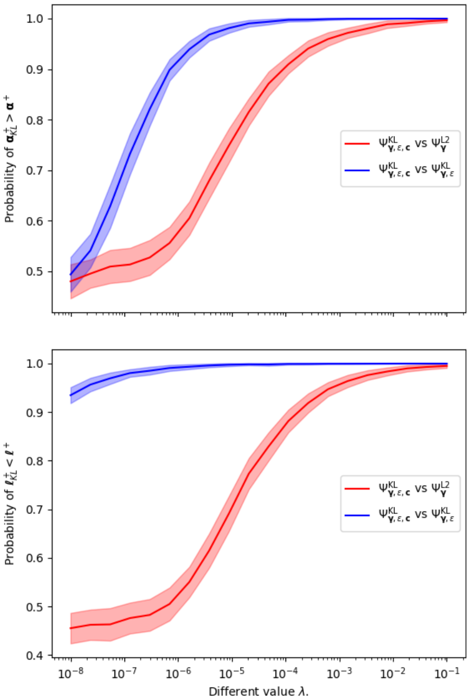


Fig. 4: Proportion of components $n \in [N]$ for which the α_n^+ (resp. ℓ_n^+ in the bottom graph) for $\Psi = \Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon,\boldsymbol{c}}$ is larger (resp. smaller) than that of $\Psi^{\mathrm{L2}}_{\boldsymbol{\gamma}}$ (red) or $\Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon}$ (blue). The curves represent the average proportions over 100 random generations of problems with size $(M, N) = (100, 256)$.

ℓ_n^+) for a given exact relaxation J_Ψ, the larger number of local (not global) local minimizers of J_0 it is likely to remove. We report in Table 1 the quantities α_n^+ and ℓ_n^+ for the generating functions considered in this work.

In Figure 4 (top) we report the proportion of components $n \in [N]$ for which the α_n^+ of the proposed $\Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon,\boldsymbol{c}}$ is larger than that of $\Psi^{\mathrm{L2}}_{\boldsymbol{\gamma}}$ (red) or $\Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon}$ (blue). Again, we repeat this for 100 problems of size $(M, N) = (100, 256)$ generated as described in [9, Section 6.3] and report the average proportion. For very small values of λ ($< 10^{-7}$) the proportion is around 50% meaning that the different relaxations behave similarly. However, when λ increases (within a range of practical interest), the reported proportions increase which illustrates the improved landscape of the proposed $\Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon,\boldsymbol{c}}$. We can draw the same conclusions for the bottom graph of Figure 4 which similarly reports the average proportion of components $n \in [N]$ for which the ℓ_n^+ of the proposed $\Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon,\boldsymbol{c}}$ is smaller than that of $\Psi^{\mathrm{L2}}_{\boldsymbol{\gamma}}$ (red) or $\Psi^{\mathrm{KL}}_{\boldsymbol{\gamma},\epsilon}$ (blue).

4. NUMERICAL SOLUTION VIA BREGMAN PROXIMAL GRADIENT ALGORITHM

A standard algorithmic choice for minimizing (4) is the Proximal Gradient Algorithm (PGA), as long as the (generally, multi-valued) proximal map of $\rho\beta_\psi$ can be computed for a given step-size $\rho > 0$. Note, however, that in order to guarantee convergence [10], the step-size should satisfy $\rho < 1/L$, with L being the Lipschitz constant of the gradient of D_{KL} with respect to $\mathbf{x}$. Following [2], an estimation of L is $L = \|\mathbf{A}\|^2/b^2$, with $b = \min_{m\in[M]} b_m$. Given that the b_m are usually small, the condition $\rho < 1/L$ leads to very small step-sizes. As a consequence, algorithmic convergence may suffer.

To overcome such practical difficulty, we employ in this section the Bregman PGA (see, e.g., [11, 12]) which naturally adapts to the Bregman-type structure of the functional J_Ψ. Compared to the standard proximal gradient step, the squared distance is here replaced by a suitable Bregman divergence and coupled with the gradient of the smooth component, in our case D_{KL}. Let us now consider $h(x) = -\log(x)$, the Burg's entropy. For all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^N_{>0}$, the Bregman divergence associated to h is given by $D_h(\mathbf{x}; \mathbf{u}) = \sum_{n=1}^N \frac{u_n}{x_n} - \log(\frac{u_n}{x_n}) - 1$. We are thus interested in the Bregman proximal operator associated to $h(\cdot)$ defined by

$$\mathrm{prox}^h_{\rho B_\Psi}(\mathbf{x}) = \underset{\mathbf{u}>0}{\mathrm{argmin}}\ B_\Psi(\mathbf{u}) + \frac{1}{\rho} D_h(\mathbf{x}, \mathbf{u}).$$

The Bregman proximal gradient step reads[1] for $k \geq 0$:

$$\mathbf{x}^{k+1} \in \mathrm{prox}^h_{\rho B_\Psi}\left(G_\rho(\mathbf{x}^k)\right) \tag{9}$$

where G_ρ is a Bregman gradient step given by:

$$G_\rho(\mathbf{x}) = \nabla H^*\left(\nabla H(\mathbf{x}) - \rho \mathbf{A}^T \nabla D_{\mathrm{KL}}(\mathbf{y}; \mathbf{A}\mathbf{x} + \mathbf{b})\right)$$

and where $H(\mathbf{x}) = \sum_{n=1}^N h(x_n)$, and H^* is its conjugate, given by $H^*(\mathbf{u}) = -\sum_{n=1}^N \log(-u_i) - N$.

Given our choice for h, the iteration (9) can thus be explicitly written as:

$$\mathbf{x}^{k+1} \in \mathrm{prox}^h_{\rho B_\Psi}\left(\frac{\mathbf{x}^k}{1 + \rho \mathbf{x}^k \mathbf{A}^T \nabla D_{\mathrm{KL}}(\mathbf{y}; \mathbf{A}\mathbf{x}^k + \mathbf{b})}\right), \tag{10}$$

where the multiplication and the division are intended element-wise. In [12, Theorem 4.1], the authors show that it suffices to take $\rho < \frac{1}{\mathcal{L}}$, where $\mathcal{L} > 0$ is the smallest constant such that the function $\mathcal{L}H(\mathbf{x}) - D_{\mathrm{KL}}(\mathbf{y}; \mathbf{A} \cdot + \mathbf{b})$ is convex ($\mathcal{L}$ is called the No-Lipschitz or the L-smad constant), to ensure convergence of this algorithm to a critical point of the objective function. In our case, we can show that $\mathcal{L} := \sum_{m=1}^M y_m$ (see [11, Section 5.2]) which no longer depends on b and $\mathbf{A}$, as opposed to the L-smoothness constant of the standard PGA.

Given the separability of B_Ψ, the computation of the Bregman proximal operator associated to $h(\cdot)$ reduces to the following problem

$$\mathrm{prox}^h_{\rho\beta_{\psi_n}}(x) = \underset{u>0}{\mathrm{argmin}} \left\{\beta_{\psi_n}(u) + \frac{1}{\rho}\left(\frac{u}{x} - \log\frac{u}{x}\right)\right\} \tag{11}$$

where we removed the constant term $-1/\rho$ in the objective. The following proposition provides a general formula for solving (11) explicitly.

[1] We present the Bregman proximal gradient iteration (9) as in [13, Section 6] to highlight connections with the standard PGA (gradient step followed by a prox step). This is equivalent to the formulation in [11, 12].

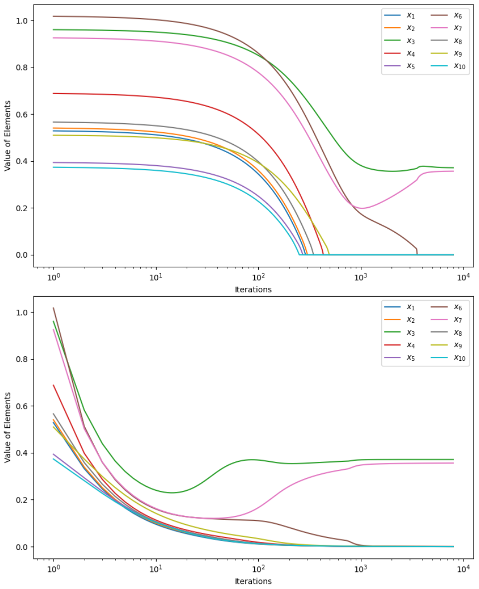


Fig. 5: Evolution of each component of the iterates along the iteration counter: PGA on the top and BPGA on the bottom for the minimization of J_Ψ (with $\Psi = \Psi^{\text{KL}}_{\boldsymbol{\gamma},\epsilon,\boldsymbol{c}}$ and $(M, N) = (5, 10)$).

Proposition 2 (Bregman proximal operator)**.** *Let $\rho > 0$, $h(x) = -\log(x)$ and $n \in [N]$. For $x > 0$, the Bregman proximal operator of $\rho\beta_{\psi_n}$ associated to $h(\cdot)$ is given by*

$$\text{prox}^h_{\rho\beta_{\psi_n}}(x) = \underset{u \in \mathcal{U}(x)}{\text{argmin}} \left\{ \beta_{\psi_n}(u) + \frac{1}{\rho}\left(\frac{u}{x} - \log\frac{u}{x}\right)\right\}$$

where $\mathcal{U}(x) := \{x\} \cup S_x$ with $S_x = \{u \in \mathbb{R}_{>0} \ : \ \frac{1}{u} + \rho\psi'_n(u) = \frac{1}{x} + \rho\psi'_n(\alpha^+_n)\}$.

Proof. By definition of the Bregman proximal operator and with the first-order optimality conditions, the possible solutions of (11) are $u^* = x$ and $u^* \neq 0$ solving $\beta'_{\psi_n}(u^*) + \frac{1}{\rho}(h'(u^*) - h(x)) = 0$, which is equivalent to $\frac{1}{u^*} + \rho\psi'_n(u^*) = \frac{1}{x} + \rho\psi'_n(\alpha^+_n)$. □

Elementary algebra yields for the choice $\psi^{\text{L2}}_{\gamma_n}(x) = \frac{\gamma_n}{2}x^2$ that the set S_x in Proposition 2 is given by:

$$S_x = \left\{ \frac{1}{2\rho\gamma_n}\left(\frac{1}{x} + \rho\gamma_n\alpha^+_n \pm \sqrt{\left(\frac{1}{x} + \rho\gamma\alpha^+_n\right)^2 - 4\rho\gamma}\right)\right\}.$$

For $\psi^{\text{KL}}_{\gamma_n,\epsilon,c_n}(x) = \gamma_n(c_n x + \epsilon - \log(c_n x + \epsilon))$, the set S_x is given by:

$$S_x = \left\{ -\frac{1}{2c_n k(x)}\left(\epsilon k(x) + \rho\gamma_n c_n - c_n \pm \sqrt{\Delta}\right)\right\},$$

where $k(x) = \frac{1}{x} - \frac{\rho\gamma_n c_n}{c_n\alpha^+_n + \epsilon}$ and $\Delta = (\epsilon k(x) + \rho\gamma_n c_n - c_n)^2 + 4c_n\epsilon k(x)$.

Ensuring a sparse solution. Under the choice of the Burg's entropy, one can observe that the Bregman proximal operator never returns zero (see Proposition 2). As such, the Bregman gradient step in (10) will never threshold any component to zero during the iterations. However, off-support components will asymptotically converge to zero, akin to multiplicative update algorithms [14, 11]. This behaviour is illustrated in Figure 5 where the evolution of each component along the iteration counter is displayed for a small-size example with $(M, N) = (5, 10)$. For this example, we observe that both PGA and BPGA converge to the same (local) minimizer of the relaxed functional J_ψ with a support of size equal to two. As a consequence of this property, we thus suggest BPGA to be initialized with a vector $\mathbf{x}^0 \in \mathbb{R}^N_{>0}$ since any initial component equal to zero will remain zero along the iterations.

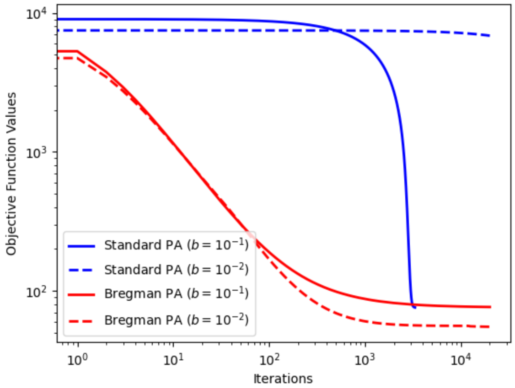


Fig. 6: Evolution of $J_\Psi(\mathbf{x}^k)$ (with $\Psi = \Psi^{\text{KL}}_{\boldsymbol{\gamma},\epsilon,\boldsymbol{c}}$) along with iterations using both the standard PGA (blue) and the Bregman PGA (red) as well as two different $\mathbf{b} = b\mathbf{1} \in \mathbb{R}^M_{>0}$ with $b \in \mathbb{R}_{>0}$.

Note that, from a practical perspective, one needs to threshold the solution provided by BPGA in order to obtain a proper sparse solution. Within our exact relaxation framework, we get from (C1) that exact relaxations J_Ψ cannot have components within $(0, \alpha^+_n)$. As such, given a solution $\hat{\mathbf{x}}$ computed by BPGA a natural thresholding rule consists in computing a sparse solution $\mathbf{x}^*$ such that $\forall n \in [N]$

$$x^*_n = \begin{cases} \hat{x}_n & \text{if } \hat{x}_n > \alpha^+_n \\ 0 & \text{otherwise} \end{cases}.$$

Convergence speed comparisons. Figure 6 illustrates the convergence of the Bregman PGA in comparison with the standard PGA for an exemplar problem of size $(M, N) = (500, 1000)$ and for two different choices of $\mathbf{b} = b\mathbf{1} \in \mathbb{R}^M_{>0}$ with $b \in \mathbb{R}_{>0}$. For both algorithms, the initial point is fixed as $\mathbf{x}^0 = \mathbf{A}^T(\mathbf{y} + \mathbf{b}) \in \mathbb{R}^N_{>0}$.

We observe the high sensitivity of PGA to the value of b. Indeed, small values of b requires the consideration of very small step-size $\rho < 1/L \propto b^2$. In contrast, the convergence of the Bregman PGA is governed by $\rho < 1/\mathcal{L}$ with $\mathcal{L}$ not depending on b. It thus achieves the same convergence speed independently on the value of this parameter. Moreover, we can also observe that BPGA is always faster than PGA.

Remark 2. *All the examples considered in this section were made of synthetic data generated as described in [9, Section 6.3]. We only reported results obtained by considering $\Psi^{\text{KL}}_{\boldsymbol{\gamma},\epsilon,\boldsymbol{c}}$. However, similar considerations can be made for other choices of Ψ.*

5. CONCLUSION

We considered continuous exact relaxations for KL ℓ_0-sparse regression problems where regularizations are defined in terms of suitable Bregman generating functions. We proposed a particular choice of the functions consistent with the geometry of the KL data term and made precise the condition required to achieve exactness of the relaxation. Differently from other generating functions, under the proposed choice, the conditions for exact relaxation can be computed explicitly and the resulting relaxation is shown to enjoy a better optimization landscape for exemplar both 1D and 2D problems.

To overcome the practical limitations in the convergence speed of the standard proximal gradient algorithm due to an over-estimation of the smoothness constant constraining the algorithmic step-size, we then considered a Bregman proximal gradient algorithm with a suitable metric for solving the relaxed problem. By accounting for tailored compatibility conditions between the algorithmic Bregman metric and the KL data term, a more robust convergence condition is found which results in more effective algorithmic performance.

6. REFERENCES

[1] M. Bertero, P. Boccacci, and V. Ruggiero, *Inverse Imaging with Poisson Data*, 2053-2563. IOP Publishing, UK, 2018.

[2] 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.

[3] T. T. Nguyen, C. Soussen, J. Idier, and E-H. Djermoune, "NP-hardness of ℓ_0 minimization problems: revision and extension to the non-negative setting," in *Proceedings of SAMPTA*, Bordeaux, 2019.

[4] E. Soubies, L. Blanc-Féraud, and G. Aubert, "A Continuous Exact ℓ_0 Penalty (CEL0) for Least Squares Regularized Problem," *SIAM Journal on Imaging Sciences*, vol. 8, no. 3, pp. 1607–1639, 2015.

[5] M. Carlsson, "On Convex Envelopes and Regularization of Non-convex Functionals Without Moving Global Minima," *Journal of Optimization Theory and Applications*, vol. 183, no. 1, pp. 66–84, 2019.

[6] Y. Liu, S. Bi, and S. Pan, "Equivalent lipschitz surrogates for zero-norm and rank optimization problems," *Journal of Global Optimization*, vol. 72, pp. 679–704, 2018.

[7] W. Bian and X. Chen, "A Smoothing Proximal Gradient Algorithm for Nonsmooth Convex Regression with Cardinality Penalty," *SIAM Journal on Numerical Analysis*, vol. 58, no. 1, pp. 858–883, 2020.

[8] M. Lazzaretti, L. Calatroni, and C. Estatico, "Weighted-CEL0 sparse regularisation for molecule localisation in super-resolution microscopy with Poisson data," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 1751–1754.

[9] M. Essafri, L. Calatroni, and E. Soubies, "Exact Continuous Relaxations of ℓ_0-Regularized Criteria with Non-quadratic Data Terms," *arXiv:2402.06483*, 2024.

[10] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods," *Mathematical Programming*, vol. 137, no. 1, pp. 91–129, 2013.

[11] H. H. Bauschke, J. Bolte, and M. Teboulle, "A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications," *Mathematics of Operations Research*, vol. 42, no. 2, pp. 330–348, 2017.

[12] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, "First Order Methods Beyond Convexity and Lipschitz Gradient Continuity with Applications to Quadratic Inverse Problems," *SIAM Journal on Optimization*, vol. 28, no. 3, pp. 2131–2151, 2018.

[13] M. Teboulle, "A simplified view of first order methods for optimization," *Mathematical Programming*, vol. 170, no. 1, pp. 67–96, 2018.

[14] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

A. PROOF OF PROPOSITION 1

Since f and ψ_n are twice differentiable, we have:

$$g''(t) = \sum_{m=1}^{M} a_{mn}^2 d''_{\mathrm{KL}}(y_m; [\mathbf{A}\mathbf{x}^{(n)}]_m + ta_{mn} + b_m) - \psi''_n(t).$$

Thus, (C1) holds if and only if for all $t \in (0, \alpha_n^+)$ there holds $g''(t) < 0$. Note that $d''_{\mathrm{KL}}(y_m; t + b_m) = \frac{y_m}{(t+b_m)^2}$, which is a decreasing function on $\mathbb{R}_{\geq 0}$. For $(\mathbf{A}, \mathbf{y}, \mathbf{x}) \in \mathbb{R}_{\geq 0}^{M\times N} \times \mathbb{R}_{\geq 0}^{M} \times \mathbb{R}_{\geq 0}^{N}$, we thus deduce that for all $m, n \in [M] \times [N]$:

$$d''_{\mathrm{KL}}(y_m; [\mathbf{A}\mathbf{x}^{(n)}]_m + ta_{mn} + b_m) \leq d''_{\mathrm{KL}}(y_m; ta_{mn} + b_m).$$

Let now $c_n = \min_{m\in\sigma(\mathbf{a}_n)} a_{mn}$ and $\epsilon \leq \min_{m\in[M]} b_m$. We have

$$a_{mn}^2 d''_{\mathrm{KL}}(y_m; a_{mn}t + b_m) \leq c_n^2 d''_{\mathrm{KL}}(y_m; c_n t + \epsilon),$$

which trivially holds for $a_{mn} = 0$ and can be proved for $a_{mn} > 0$ by observing:

$$a_{mn}^2 d''_{\mathrm{KL}}(y_m; a_{mn}t + b_m) = \frac{a_{mn}^2 y_m}{(a_{mn}t + b_m)^2} \leq \frac{a_{mn}^2 y_m}{(c_n t + \epsilon)^2}$$

Finally, we thus get

$$\begin{aligned} g''(t) &\leq \sum_{m=1}^{M} a_{mn}^2 d''_{\mathrm{KL}}(y_m; a_{mn}t + b_m) - \psi''_n(t) \\ &\leq \sum_{m=1}^{M} \frac{a_{mn}^2 y_m}{(c_n t + \epsilon)^2} - \psi''_n(t). \end{aligned} \tag{12}$$

Let us consider now $\psi_n(t) = \psi^{\mathrm{KL}}_{\gamma_n,\epsilon,c_n}(t) = \gamma_n(c_n t + \epsilon - \log(c_n t + \epsilon))$ for all $n \in [N]$. We have $\psi''_n(t) = \gamma_n \frac{c_n^2}{(c_n t+\epsilon)^2}$. Thus, (12) simplifies to

$$g''(t) \leq \sum_{m=1}^{M} \frac{a_{mn}^2 y_m}{(c_n t + \epsilon)^2} - \gamma_n \frac{c_n^2}{(c_n t + \epsilon)^2} < 0,$$

which implies

$$\sum_{m=1}^{M} \frac{a_{mn}^2 y_m}{c_n^2} < \gamma_n.$$

which completes the proof.