Box-constrained ℓ_0 Bregman relaxations

Mhamed Essafri
IRIT, Université de Toulouse, INP
Toulouse, France
mhamed.essafri@irit.fr

Luca Calatroni
MaLGa, DIBRIS, Università di Genova,
MMS, Istituto Italiano di Tecnologia,
Genoa, Italy,
luca.calatroni@unige.it

Emmanuel Soubies
IRIT, Université de Toulouse, CNRS
Toulouse, France
emmanuel.soubies@cnrs.fr

***Abstract*—Regularization using the ℓ_0 pseudo-norm is a common approach to promote sparsity, with widespread applications in machine learning and signal processing. However, solving such problems is known to be NP-hard. Recently, the ℓ_0 Bregman relaxation (B-rex) has been introduced as a continuous, non-convex approximation of the ℓ_0 pseudo-norm. Replacing the ℓ_0 term with B-rex leads to exact continuous relaxations that preserve the global optimum while simplifying the optimization landscape, making non-convex problems more tractable for algorithmic approaches. In this paper, we focus on box-constrained exact continuous Bregman relaxations of ℓ_0-regularized criteria with general data terms, including least-squares, logistic regression, and Kullback-Leibler fidelities. Experimental results on synthetic data, compared with Branch-and-Bound methods, demonstrate the effectiveness of the proposed relaxations.**

***Index Terms*—ℓ_0-relaxation, non-convex optimization, continuous exact relaxations, box constraint.**

I. Introduction

Given a possibly undetermined ($M \leq N$) forward matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ and vector of observations $\mathbf{y} \in \mathbb{R}^M$, we consider problems of the form

$$\hat{\mathbf{x}} \in \underset{\mathbf{x}\in[l,u]^N}{\operatorname{argmin}} \left\{ J_0(\mathbf{x}) := F_{\mathbf{y}}(\mathbf{A}\mathbf{x}) + \lambda_0 \|\mathbf{x}\|_0 + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2 \right\} \quad (1)$$

where $l \in \mathbb{R}_{\leq 0} \cup \{-\infty\}$ and $u \in \mathbb{R}_{\geq 0} \cup \{+\infty\}$ define a box constraint, the terms $\|\cdot\|_2$ and $\|\cdot\|_0$ denotes respectively the squared ℓ_2 norm and the ℓ_0 pseudo-norm that counts the number of non-zero elements in a vector of $\mathbb{R}^N$. The hyperparameters $\lambda_0 > 0$ and $\lambda_2 \geq 0$ control respectively the sparsity and the ℓ_2 ridge regularization strengths. Finally, $F_{\mathbf{y}} : \mathbb{R}^M \mapsto \mathbb{R}_{\geq 0}$ is a data-fidelity function that measures the discrepancy between the model $\mathbf{A}\mathbf{x}$ and the data $\mathbf{y}$, and satisfies the following assumption.

Assumption 1: The data fidelity function is coordinate-wise separable. i.e, $F_{\mathbf{y}}(\mathbf{z}) = \sum_{m=1}^M f(z_m; y_m)$, where for each $y \in \mathbb{R}$, $f(\cdot; y)$ is convex, proper, twice differentiable on (l, u) and bounded from below.

Beyond the popular least-squares function, exemplar instances of such data-fidelity terms include the Kullback-Leibler divergence [1] and logistic loss [2], which arise in signal/image processing and machine learning applications.

ME and ES acknowledge the financial support of the ANR EROSION (ANR-22-CE48-0004). LC acknowledges the financial support of the European Research Council (grant MALIN, 101117133).

Remark 1: To simplify the presentation, we consider in (1) the case where $l := l_1 = l_2 = \cdots = l_N$ and $u := u_1 = u_2 = \cdots = u_N$, i.e. the same box constraint is applied to each component. Note, however, that our results can be easily extended to the general case.

A. Related Works

While the ℓ_0 pseudo-norm is the most natural choice for enforcing sparsity, its discontinuity, non-convexity, and non-smoothness make the associated problem (1) NP-hard [3]. Yet there exists a vast literature related to this problem. One approach is to address the original problem directly, such as with the Iterative Hard Thresholding (IHT), which can be extended to this setting and guarantees convergence to a critical point [4]. In [5], the authors studied the proximal mapping of the ℓ_0 function over symmetric sets that satisfy a submodularity-like property (SOM) and developed algorithms that converge to critical points. For moderately sized problems, branch-and-bound (BnB) methods offer exact solutions at a reasonable computational cost [6], [7].

In this work, we focus on exact relaxation approaches, which replace the ℓ_0 pseudo-norm with a continuous (non-convex) penalty function while preserving global minimizers [8]–[13]. Moreover, such relaxed formulations remove some local minimizers of the initial problem, making the optimization landscape more favorable to optimization algorithms. In this context, the authors in [8] proposed an exact relaxation using a capped-ℓ_1 penalty for cases where $F_{\mathbf{y}}(\mathbf{A}\cdot) + \frac{\lambda_2}{2}\|\cdot\|_2^2$ is convex, Lipschitz continuous, and non-smooth, applicable to both unconstrained and box-constrained cases. Building on this, they developed a smoothing proximal gradient (SPG) algorithm to find a stationary point of the relaxed problem, which corresponds to a local minimizer of the original problem. However, for the previously mentioned data-terms, $F_{\mathbf{y}}(\mathbf{A}\cdot) + \frac{\lambda_2}{2}\|\cdot\|_2^2$ is not necessarily Lipschitz continuous. For this setting, quadratic envelopes of the ℓ_0 pseudo-norm [9], including the continuous exact ℓ_0 (CEL0) penalty [10], enable exact relaxation for least-squares data terms. Recently, these ideas have been generalized with the introduction of the ℓ_0 Bregman relaxation (B-rex), providing exact relaxations for general (non-quadratic) data terms [11], [12]. Additionally, the work in [13] proposed a weighted-CEL0 relaxation for weighted-ℓ_2 data terms, offering an approximation of the KL divergence. While these approaches are limited to the

unconstrained case, in this paper, we extend the works of [11], [12] to box-constrained problems.

B. Contributions and Outline

In this work we extend the framework of exact continuous ℓ_0 Bregman relaxations, originally proposed for unconstrained and non-negatively constrained problems [11], [12], to box-constrained problems of the form (1). More precisely, we provide in Proposition 2 an exact relaxation result. In view of designing effective numerical schemes solving the relaxed problem, we consider in Section IV some proximal-based schemes showing, in particular, how the proximal operator of the proposed exact penalty can be efficiently computed. In addition, we will show that the relaxed problem can be minimized via iteratively reweighted ℓ_1. In Section V, we compare the minimization of the relaxed criteria with both BnB and IHT procedures showing good agreement with certified global procedures in the case of small-size problems and the applicability of the approach to larger-scale problems.

C. Notations

Let $[N] = \{1, 2, \dots, N\}$ denote the set of indices up to N. The symbol $\mathbf{0} \in \mathbb{R}^N$ represents the vector of all zeros, and for each $n \in [N]$, $\mathbf{e}_n \in \mathbb{R}^N$ denotes the unit vector of the standard basis in $\mathbb{R}^N$. The set $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} \mid x \geq 0\}$ represents the non-negative real line. For $n \in [N]$, $\mathbf{x}^{(n)} = (x_1, \dots, x_{n-1}, 0, x_{n+1}, \dots, x_N)$ denotes the vector $\mathbf{x}$ with the n-th component replaced by 0.

II. Box-Constrained B-rex

In this section, we extend the ℓ_0 Bregman Relaxation (B-rex) introduced in [11] to the box-constrained Problem (1). Consider a family $\Psi = \{\psi_n\}_{n\in[N]}$ of strictly convex, proper, and twice-differentiable functions such that $\mathrm{dom}(\psi_n) \in \{\mathbb{R}, \mathbb{R}_{\geq 0}\}$ (note that $\mathbb{R}_{\geq 0}$ is allowed only when $l = 0$). We define the box-constrained B-rex as $B^{l,u}_\Psi : [l, u]^N \to \mathbb{R}_{\geq 0}$ such that for all $\mathbf{x} \in [l, u]^N$:

$$B^{l,u}_\Psi(\mathbf{x}) = \sup_{(\alpha,\mathbf{z})\in\mathcal{C}} \alpha - D_\Psi(\mathbf{x}, \mathbf{z}) \tag{2}$$

where

$$\mathcal{C} = \left\{(\alpha, \mathbf{z}) \in \mathbb{R} \times \mathrm{dom}(\Psi) \text{ s.t. } \alpha - D_\Psi(\cdot, \mathbf{z}) \leq \lambda_0 \|\cdot\|_0 + \mathcal{I}_{[l,u]^N}\right\}$$

and $\mathcal{I}_{[l,u]^N}(\mathbf{x}) = \{0 \text{ if } \mathbf{x} \in [l, u]^N; +\infty \text{ otherwise}\}$ and D_Ψ denotes the Bregman divergence associated to Ψ. It is defined, for all $\mathbf{x}, \mathbf{z} \in \mathrm{dom}(\Psi)$ by

$$D_\Psi(\mathbf{x}, \mathbf{z}) = \sum_{n=1}^{N} d_{\psi_n}(x_n, z_n) \tag{3}$$

with $d_{\psi_n}(x, z) = \psi_n(x) - \psi_n(z) - \psi'_n(z)(x - z)\ \forall x, z \in \mathrm{dom}(\psi_n)$. Standard choices of the functions ψ_n are the p-power functions, the Shannon entropy and the Kullback-Leibler divergence, see [11], [12]. Note, that as opposed to the unconstrained B-rex proposed therein, the extension (2) depends on the box-constraint through the term $\mathcal{I}_{[l,u]^N}$.

Proposition 1 (Closed form expression of $B^{l,u}_\Psi$): For all $n \in [N]$, let $\alpha_n^- \leq 0$ and $\alpha_n^+ \geq 0$ be such that $[\alpha_n^-, \alpha_n^+]$ defines the λ_0-sublevel set of $d_{\psi_n}(0, \cdot)$. Then, for every $\mathbf{x} \in [l, u]^N$, we have $B^{l,u}_\Psi(\mathbf{x}) = \sum_{n=1}^N \beta^{l,u}_{\psi_n}(x_n)$, where, for $x \in [l, u]$, the functions $\beta^{l,u}_{\psi_n}$ are defined by

$$\beta^{l,u}_{\psi_n}(x) = \begin{cases} \psi_n(0) - \psi_n(x) + \kappa_n^- x, \text{ if } x \in (\eta_n^-, 0], \\ \psi_n(0) - \psi_n(x) + \kappa_n^+ x, \text{ if } x \in [0, \eta_n^+), \\ \lambda_0, \qquad\qquad\qquad\quad \text{if } x \in [l, u]\backslash(\eta_n^-, \eta_n^+). \end{cases}$$

where $\eta_n^- = \max\{\alpha_n^-, l\}, \quad \eta_n^+ = \min\{\alpha_n^+, u\}$. Moreover, for $l \neq 0$ and $u \neq 0$, the slopes κ_n^- and κ_n^+ are given by

$$\kappa_n^- = \begin{cases} \psi'_n(\alpha_n^-), & \text{if } \alpha_n^- \geq l, \\ l^{-1}\left(\lambda_0 + \psi_n(l) - \psi_n(0)\right), & \text{if } \alpha_n^- < l, \end{cases}$$

$$\kappa_n^+ = \begin{cases} \psi'_n(\alpha_n^+), & \text{if } \alpha_n^+ \leq u, \\ u^{-1}\left(\lambda_0 + \psi_n(u) - \psi_n(0)\right), & \text{if } \alpha_n^+ > u. \end{cases}$$

The proof can be found in the supplementary material of this paper available online [14]. In Figure 1, we present one-dimensional examples of box-constrained B-rex. We distinguish two cases. In the first one (left graph), the box-constraint $[l, u]$ contains the interval $[\alpha^-, \alpha^+]$ where the unconstrained B-rex [11] is non-constant. In this case, both box-constrained and unconstrained B-rex coincide. In the opposite situation where $[l, u] \subset [\alpha^-, \alpha^+]$ (right graph), the box-constrained B-rex deviates from its unconstrained (showed in gray) counterpart.

Equipped with the box-constrained B-rex (2), we consider the following continuous relaxation of J_0 for $\mathbf{x} \in [l, u]^N$,

$$J^{l,u}_\Psi(\mathbf{x}) = F_\mathbf{y}(\mathbf{A}\mathbf{x}) + B^{l,u}_\Psi(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2. \tag{4}$$

III. Exact Relaxations Properties

In Theorem 2, we provide conditions on Ψ so that J_Ψ is an *exact relaxation* of J_0, as originally defined in [10], meaning that it preserves global minimizers while eliminating certain local ones.

Theorem 2 (Exact relaxation property): Let Ψ be such that, $\forall n \in [N]$ and $\forall t \in (\eta_n^-, 0) \cup (0, \eta_n^+)$

$$\frac{\partial^2}{\partial t^2} F_\mathbf{y}(\mathbf{A}(\mathbf{x}^{(n)} + t\mathbf{e}_n)) + \lambda_2 < \psi''_n(t), \tag{5}$$

where $\mathbf{x}^{(n)} = (x_1, \dots, x_{n-1}, 0, x_{n+1}, \dots, x_N)^T$. Then,

$$\operatorname*{argmin}_{\mathbf{x}\in[l,u]^N} J^{l,u}_\Psi(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{x}\in[l,u]^N} J_0(\mathbf{x}), \tag{6}$$

$$\hat{\mathbf{x}} \text{ local minimizer of } J^{l,u}_\Psi \text{ over } [l, u]^N \implies \hat{\mathbf{x}} \text{ local minimizer of } J_0 \text{ over } [l, u]^N. \tag{7}$$

Proof. This result is a direct generalization of [11, Theorem 9]. It relies on three facts: i) $J^{l,u}_\Psi(\mathbf{x}) \leq J_0(\mathbf{x})\ \forall \mathbf{x} \in [l, u]^N$ (by definition of $B^{l,u}_\Psi$), ii) $\forall \mathbf{x} \notin \prod_{n\in[N]}(\eta_n^-, 0) \cup (0, \eta_n^+)$, $J^{l,u}_\Psi(\mathbf{x}) = J_0(\mathbf{x})$ (from Proposition 1) and iii) under (5) $\forall n \in [N]$, $t \mapsto J^{l,u}_\Psi(\mathbf{x}^{(n)} + t\mathbf{e}_n)$ is strictly concave on both $(\eta_n^-, 0)$ and $(0, \eta_n^+)$. □

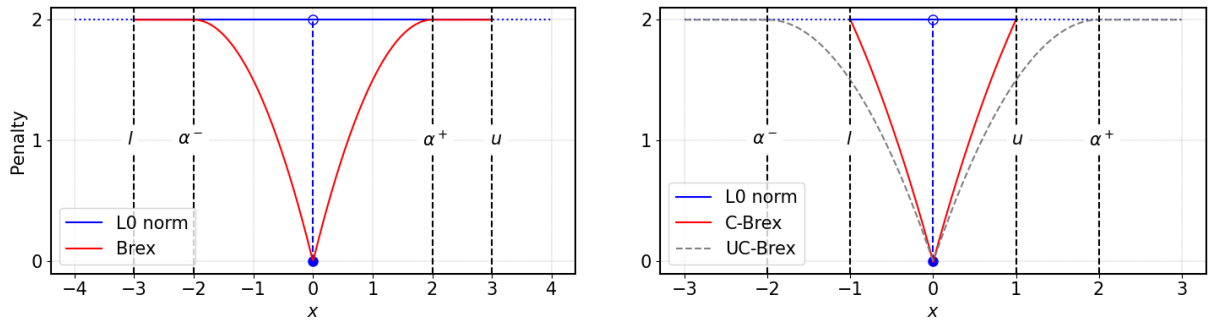

Fig. 1: Box-constrained B-rex ($\lambda_0 = 2$, $\psi = 1/2\|\cdot\|_2^2$) when (left) $[l,u] \supseteq [\alpha^-, \alpha^+]$ and (right) $[l,u] \subseteq [\alpha^-, \alpha^+]$.

IV. Minimizing $J_\Psi^{l,u}$

In this section, we discuss the optimization of the relaxed problem through two different optimization algorithms: the forward backward splitting (FBS) [15] and the iteratively reweighted ℓ_1 (IRL1) [16].

A. Forward-Backward Splitting Algorithm

The FBS or proximal gradient algorithm consists of the following time stepping scheme for a given initialisation $\mathbf{x}^0$ and step-size $\rho > 0$:

$$\mathbf{x}^{k+1} \in \operatorname{prox}_{\rho(B_\Psi^{l,u} + \mathcal{I}_{[l,u]^N})} \left(\mathbf{x}^k - \rho(\mathbf{A}^T \nabla F_{\mathbf{y}}(\mathbf{A}\mathbf{x}^k) + \lambda_2 \mathbf{x}^k)\right).$$

It requires the ability to efficiently evaluate the proximal operator of $B_\Psi^{l,u} + \mathcal{I}_{[l,u]^N}$. Since B-rex is a separable penalty, this computation reduces to evaluate independent one-dimensional proximal operators, for which a closed-form expression can be derived (for standard choices of Ψ) from the following proposition.

Proposition 3: Let $\rho > 0$ and $n \in [N]$. For $x \in \mathbb{R}$, the proximal operator of $\rho\beta_{\psi_n}^{l,u}$ is given by

$$\operatorname{prox}_{\rho\beta_{\psi_n}^{l,u}}(x) = \operatorname*{argmin}_{v \in [l,u] \cap \mathcal{V}(x)} \left\{ \beta_{\psi_n}^{l,u}(v) + \frac{1}{2\rho}(v-x)^2 \right\} \quad (8)$$

where $\mathcal{V}(x) = \{l, 0, x, u\} \cup S_x$ with $S_x = \{v \in \mathbb{R} : v - \rho\psi_n'(v) = x - \rho\kappa_n^{\pm}\}$.

Proof. The proof can be found in Appendix A. □

The functional $J_\Psi^{l,u}$ satisfies the Kurdyka-Łojasiewicz property, hence taking $0 < \rho < 1/L$, with L being the Lipschtiz constant of the gradient of $F_{\mathbf{y}}(\mathbf{A}\cdot) + \lambda_2/2\|\cdot\|_2^2$, the sequence $\{\mathbf{x}^k\}_k$ generated by FBS converge to a critical point of $J_\Psi^{l,u}$ [4]. Alternatively, a backtracking strategy can be used (see Section V) to estimate the step-size at each iteration, which helps improving the convergence speed.

B. Iteratively Reweighted ℓ_1

The Iteratively Reweighted ℓ_1 (IRL1) algorithm belongs to the class of majorization-minimization (MM) algorithms, which iteratively construct minimizing sequences of surrogate functions that upper-bound the original objective function. The optimization process consists of two main steps. First, in the *majorization step*, the objective function is upper-bounded by a surrogate equal to it at the current point. For symmetric B-rex (i.e., $\beta_{\psi_n}^{l,u}(x) = \beta_{\psi_n}^{l,u}(|x|)\ \forall (x,n) \in [l,u] \times [N]$, as those we use in our experiments), we consider the following weighted ℓ_1-norm

$$\tilde{J}(\mathbf{x}) = F_{\mathbf{y}}(\mathbf{A}\mathbf{x}) + \sum_{n=1}^{N} w_n |x_n| + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2,$$

where the weights $\{w_n\}_n$ are such that $w_n \in \partial\beta_{\psi_n}^{l,u}(|x_n|)$. Then, the *minimization step* consists of minimizing $\tilde{J}$ over the set $[l,u]^N$. This is achieved using the Projected FBS algorithm by simply considering the (closed-form) proximal operator of $\mathbf{x} \mapsto \sum_{n=1}^N w_n|x_n| + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2$. Again as the Kurdyka- Łojasiewicz property holds for $J_\Psi^{l,u}$, convergence of the iterative scheme to a critical point can be shown following [16].

V. Experiments

A. Data Generation

Forward Matrix Generation: We generate a the matrix $\mathbf{A} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ of size 500×1000 from a multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma}$. The covariance $\mathbf{\Sigma}$ follows an exponential correlation model, where its entries are defined by $\sigma_{ij} = \rho^{|i-j|}$ for $1 \le i,j \le N$, with the parameter $\rho \in [0,1]$ controlling the correlation strength. We then consider two differents settings to generate the observation vector $\mathbf{y} \in \mathbb{R}^M$ according to the data term $F_{\mathbf{y}}$ of interest.

- *Least-Square (LS)*, $F_{\mathbf{y}}(\mathbf{A}\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2$: following [7], we define a sparse vector $\mathbf{x}^* \in \mathbb{R}^N$ with $k^* \in \mathbb{N}$ non-zero equispaced entries, each sampled from a uniform distribution in the interval $[l,u]$. The observation $\mathbf{y}$ is then generated by $\mathbf{y} = \mathbf{A}\mathbf{x}^* + \boldsymbol{\varepsilon}$, where $\varepsilon_i \sim N(0, \sigma^2)$. The signal-to-noise ratio (SNR) measure defined as
$$\mathrm{SNR} = \frac{\mathrm{Var}(\mathbf{A}\mathbf{x}^*)}{\mathrm{Var}(\boldsymbol{\varepsilon})} = \frac{(\mathbf{x}^*)^T \mathbf{\Sigma} \mathbf{x}^*}{\sigma^2}$$
is used to control the level of noise on the data $\mathbf{y}$.
- *Logistic Regression (LR)*, $F_{\mathbf{y}}(\mathbf{A}\mathbf{x}) = \sum_{m=1}^M \log(1 + \exp([\mathbf{A}\mathbf{x}]_m)) - y_m[\mathbf{A}\mathbf{x}]_m$: As for LS, we generate a k^*-sparse vector $\mathbf{x}^*$ with equispaced non-zero entries, each equal to 1. Each coordinate of

the label vector $\mathbf{y}$ is binary, with $y_m \in \{-1, 1\}$, where $y_m = 1$ is determined with probability

$$P\left(y_m = 1 \mid \mathbf{a}_m\right) = 1/\left(1 + \mathbf{e}^{-s\langle \mathbf{a}_m, \mathbf{x}^* \rangle}\right),$$

where $\mathbf{a}_m$ is the m-th row of $\mathbf{A}$ and $s > 0$ controls the SNR. The labels $\mathbf{y}$ are sampled from a Bernoulli distribution.

B. Algorithmic Setting and Parameter Selection

Benchmarked Algorithms: We compare the performance of the following four methods in solving Problem (1) for 20 realizations of forward matrices $\mathbf{A}$ and vectors $\mathbf{y}$.

- *IHT* [4] on the original problem,
- *FBS* on the proposed exact relaxation (Section IV-A)
- *IRL1* on the proposed exact relaxation (Section IV-B)
- *BnB* [7] on the original problem. It can guarantee the convergence to a global minimizer for small-scale problems.

Parameters: For LS problems, initialization was set as $\mathbf{x}_0 = \mathbf{0}$ for all algorithms. The same choice was also made for LR problems for all algorithms except for IHT, where the initialization $\mathbf{x}_0 = \mathbf{A}^T\mathbf{y}$ was considered. Indeed, as shown in [11, Lemma 1], $\mathbf{x}_0 = \mathbf{0}$ is a local minimizer of J_0 (note that $\mathbf{0}$ is, in contrast and algorithms that minimize J_0 (as IHT) may thus be stuck at the initial point $\mathbf{x}_0 = \mathbf{0}$. As for FBS, we implemented IHT with a backtracking strategy to accelerate convergence and to favour (upon large initial time steps) escape from the local minimizer $\mathbf{x}_0 = \mathbf{0}$ for the LS case. As this turned out to be harder for LR, a different initialization was chosen instead. We fixed the convergence tolerance to a relative change between consecutive iterates below 10^{-7}.

The hyperparameter λ_0 is set as $\lambda_0 = \alpha F_{\mathbf{y}}(\mathbf{0})$, with $\alpha \in (0, 1)$ chosen so that the BnB solver provides a solution with a support cardinality close to the ideal k^*, in all experiments. For the LS problem, we set $\lambda_2 = 0$, focusing purely on sparsity, while for the LR problem we used $\lambda_2 = 1$. As far as the generating functions ψ_n are considered, we set $\psi_n(x) = (\gamma_n/2)x^2$, with $\gamma_n = \lambda_2 + \|\mathbf{a}_n\|_2^2 + 1e^{-10} > \lambda_2 + \|\mathbf{a}_n\|_2^2$ for LS and $\gamma_n = \lambda_2 + 0.25\|\mathbf{a}_n\|_2^2$ for LR. This choice ensures that the exact relaxation condition (5) is satisfied.

C. Results and Discussion

We assess the quality of the minimization using the value of the original function J_0 at convergence. For each algorithm, we thus computed this value for each realization of $(\mathbf{A}, \mathbf{y})$.

BnB with certification of the global solution: The left panels of Figures 2 and 3 present results for LS and LR problems when $k^* = 10$ and $k^* = 7$, respectively. In these cases, the BnB solver consistently finds (and certifies) the global optimum. From these results we can thus observe that the minimization of the proposed exact relaxation with either IRL1 or FBS also often leads to the global minimizer. In contrast, this is not the case when directly tackling the original problem with IHT.

BnB without certified solutions: In the right panels of Figures 2 and 3, we consider the case of less sparse vectors ($k^* = 25$), where BnB fails to certify optimality within the given time limit (30 minutes, whereas FBS and IRL1 solve the relaxed problem in an average time of 2.9s/3.87s for the LS cases and 0.63s/0.62s for the LR cases, respectively). In the LS setting, BnB appears to be far from the global optimum, as minimizing the relaxed problem yields solutions with a lower objective function value. Conversely, in the LR setting, while BnB does not certify optimality, it still finds solutions with the lowest objective function value, followed by methods minimizing the proposed relaxations. The difference in performance of the BnB between LS and LR can be attributed to the presence of the ℓ_2 term in the case of LR ($\lambda_2 > 0$). As shown in [7], the incorporation of the ℓ_2 term improves the relaxation quality for pruning tests, which explains why BnB performs better in the LR setting. This is further supported by the additional experiment we provide in the supplementary material [14]. There, we set $\lambda_2 > 0$ in the LS experiment with $k^* = 25$ and one can observe that the BnB performs better than the other methods.

VI. Conclusion

We introduced the constrained ℓ_0 Bregman relaxation which continuously approximate the ℓ_0 pseudo-norm within a bounded domain $[l, u]^N$. By replacing the ℓ_0 pseudo-norm with this constrained B-rex, a continuous optimization problem that preserves the same global minimizers as the original one and exhibit fewer local minimizers is obtained. This relaxation makes the problem more amenable to standard non-convex optimization algorithms, such as the proximal gradient method and the iteratively reweighted ℓ_1 (IRL1) algorithms. Through several numerical experiments, we demonstrated that our approach compares well to Branch-and-Bound (BnB) approaches, achieving strong agreement with certified global solutions provided by BnB in small-scale problem settings.

Appendix

A. Proof of Proposition 3

The proof follows from the fact that (see [11, Section 3.2]):

$$\partial\beta_{\psi_n}^{l,u}(v) = \begin{cases} \{-\psi'(v) + \kappa_n^-\} & \text{if } v < 0 \\ \{-\psi'(v) + \kappa_n^+\} & \text{if } v > 0 \\ \{-\psi'(0)\} + [\kappa_n^-, \kappa_n^+] & \text{if } v = 0 \end{cases}$$

Recalling that

$$\operatorname{prox}_{\rho\beta_{\psi_n}^{l,u}}(x) = \operatorname*{argmin}_{v \in [l,u]} \left\{ \beta_{\psi_n}^{l,u}(v) + \frac{1}{2\rho}(v - x)^2 \right\},$$

the first-order optimality condition states that

$$0 \in \frac{1}{\rho}(v - x) + \partial\beta_{\psi_n}^{l,u}(v) + \mathcal{N}_{[l,u]}(v),$$

where $\mathcal{N}_{[l,u]}$ is the normal cone[1] to $[l, u]$, defined as

$$\mathcal{N}_{[l,u]}(v) = \begin{cases} \{0\} & \text{if } v \in (l, u), \\ t \geq 0 & \text{if } v = u, \\ t \leq 0 & \text{if } v = l. \end{cases}$$

[1]The normal cone of a set $\mathcal{C} \subset \mathbb{R}$ is defined as: $\forall x \in \mathcal{C},\ \mathcal{N}_{\mathcal{C}}(x) = \{t \in \mathbb{R} \mid \langle t, z - x \rangle \leq 0\ \forall z \in \mathcal{C}\}$.

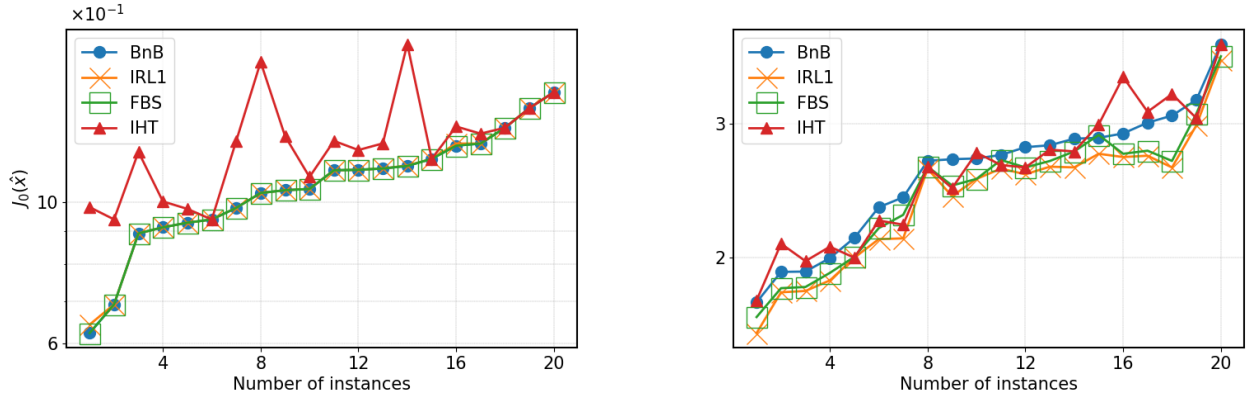

Fig. 2: LS: (ordered) values $J_0(\hat{\mathbf{x}})$ obtained by each method along the 20 problem instances. For the left plot $\lambda_0 = 2\times10^{-2}F_{\mathbf{y}}(\mathbf{0})$ and $k^* = 10$. For the right plot $\lambda_0 = 5 \times 10^{-3}F_{\mathbf{y}}(\mathbf{0})$ and $k^* = 25$. Finally, $[l, u] = [-1.5, 1.5]$, $\rho = 0.9$ and $\mathrm{SNR} = 10$.

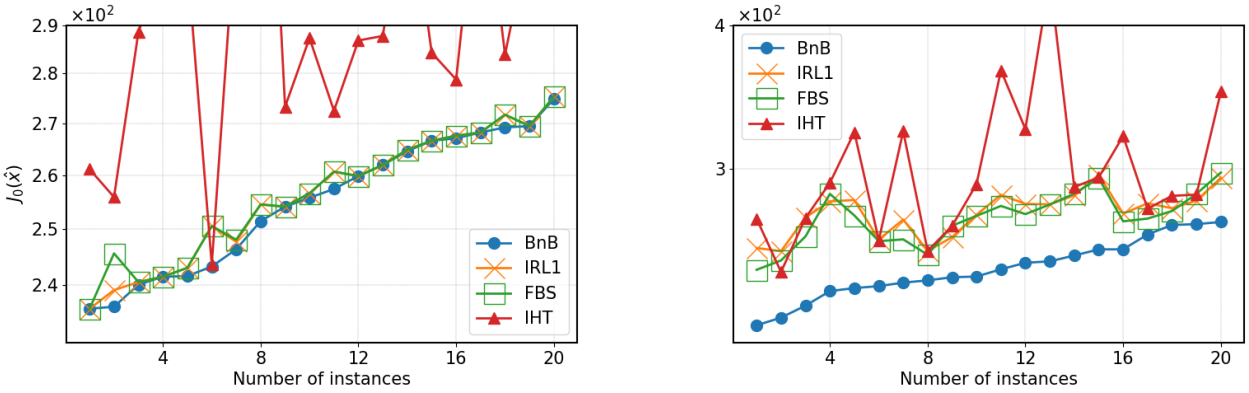

Fig. 3: LR: (ordered) values $J_0(\hat{\mathbf{x}})$ obtained by each method along the 20 problem instances. For the left plot $\lambda_0 = 2.5 \times 10^{-2}F_{\mathbf{y}}(\mathbf{0})$ and $k^* = 7$. For the right $\lambda_0 = 1.5 \times 10^{-2}F_{\mathbf{y}}(\mathbf{0})$ and $k^* = 25$. Finally $[l, u] = [-1, 1]$, $\lambda_2 = 1$, $\rho = 0.9$, $s = 1$.

It follows that the possible solutions of the proximal operator are included in $\{0, x, l, u\}\cup S_x$, where S_x is the set of solutions of the equation: $-\psi_n'(v) + \kappa_n^{\pm} + \frac{1}{\rho}(v - x) = 0$.

References

[1] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*. Cambridge university press, 2013, no. 53.

[2] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[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] 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.

[5] A. Beck and N. Hallak, "Proximal mapping for symmetric penalty and sparsity," *SIAM Journal on Optimization*, vol. 28, no. 1, pp. 496–527, 2018.

[6] D. D. Donne, M. Kowalski, and L. Liberti, "A novel integer linear programming approach for global l0 minimization," *Journal of Machine Learning Research*, vol. 24, no. 382, pp. 1–28, 2023.

[7] T. Guyard, C. Herzet, C. Elvira, and A.-N. Arslan, "A new branch-and-bound pruning framework for l0-regularized problems," in *International Conference on Machine Learning (ICML)*. PMLR, 2024.

[8] 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.

[9] 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.

[10] 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.

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

[12] ——, "On ℓ_0 Bregman-Relaxations for Kullback-Leibler Sparse Regression," in *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2024, pp. 1–6.

[13] 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.

[14] "Supplementary material," https://www.irit.fr/~Emmanuel.Soubies/Papiers/SuppMaterialBoxBrex.pdf.

[15] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[16] P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock, "On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision," *SIAM Journal on Imaging Sciences*, vol. 8, no. 1, pp. 331–372, 2015.