A Modified Exact Penalty Approach for General Constrained ℓ_0-Sparse Optimization Problems

Christian Kanzow* Felix Weiß†

September 2, 2025

Abstract. We consider a general class of constrained optimization problems with an additional ℓ_0-sparsity term in the objective function. Based on a recent reformulation of this difficult ℓ_0-term, we consider a nonsmooth penalty approach which differs from the authors' previous work by the fact that it can be directly applied to problems which do not necessarily contain nonnegativity constraints. This avoids a splitting of free variables into their positive and negative parts, reduces the dimension and fully exploits the one-to-one correspondence between local and global minima of the given ℓ_0-sparse optimization problem and its reformulation. The penalty approach is shown to be exact in terms of minima and stationary points. Since the penalty function is (mildly) nonsmooth, we also present practical techniques for the solution of the subproblems arising within the penalty formulation. Finally, the results of an extensive numerical testing are provided.

Keywords. Sparse optimization; global minima; local minima; stationary points; exact penalty function

Mathematical Science Classification. 65K05, 90C26, 90C30, 90C46

1 Introduction

In this paper, we consider the sparse optimization problem of the form

$$\min_x f(x) + \rho \|x\|_0, \quad g(x) \in X, \tag{SPO}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ and $g\colon \mathbb{R}^n \to \mathbb{R}^m$ are smooth functions, $X \subseteq \mathbb{R}^n$ is a nonempty and closed set, $\rho > 0$ a given scalar, and

$$\|x\|_0 := \text{number of nonzero components } x_i \text{ of } x.$$

Solving the (SPO) is known to be a difficult task, due to the combinatorial nature of the optimization problem induced by the nonconvex and discontinuous ℓ_0-norm. It was shown in [26], that (SPO) belongs to the class of NP-hard problems, when paired with a quadratic function f.

*University of Würzburg, Institute of Mathematics, Campus Hubland Nord, Emil-Fischer-Str. 30, 97074 Würzburg, Germany; christian.kanzow@uni-wuerzburg.de

†University of Würzburg, Institute of Mathematics, Campus Hubland Nord, Emil-Fischer-Str. 30, 97074 Würzburg, Germany; felix.weiss@uni-wuerzburg.de

The sparse optimization problem is closely related to the cardinality constrained optimization problem, where the ℓ_0-term in the objective function is replaced by a constraint of the form $\|x\|_0 \leq \kappa$. In the survey article [30], it was shown that any solution of (SPO) solves the corresponding cardinality constrained problem with $\kappa = \|x^*\|$. However, in general, it is not true that a solution x^* of the cardinality constrained problems solves (SPO) for some suitable ρ. Hence, we cannot substitute the program (SPO) by a suitable cardinality constrained optimization problem.

Following [19], solution strategies for ℓ_0-norm regularized problems can be divided into the following categories: (a) convex approximations, (b) nonconvex approximations, and (c) nonconvex exact reformulations. The first approach typically replaces the ℓ_0-term by the convex ℓ_1-norm and is particularly popular for LASSO-type problems, cf. [29]. On the other hand, in many applications, this technique does not yield the desired sparsity, and sometimes it is not even of any help in reducing the sparsity like for portfolio optimization problems, cf. [17]. The techniques based on (b) try to approximate the ℓ_0-term by a suitable (smoother) function. Popular realizations of this idea use the ℓ_p-quasi norm for $p \in (0, 1)$, the SCAD (= smoothly clipped absolute deviation) [12], MCP (= minimax concave penalty) [32], PiE (= piecewise exponential) [25], or the transformed ℓ_1 function [33], see also the list of functions in [8]. Note that all these approximations of the ℓ_0-quasi norm are at least continuous, most of them even Lipschitz continuous.

Regarding category (c), there exist different exact reformulations of problem (SPO). Assuming that there are finite lower and upper bounds as part of the standand constraints $g(x) \in X$, it is straightforward to obtain a mixed-integer formulation of the original problem, see [3, 30]. Hence, for a convex quadratic f and linear constraints, mixed-integer QP solvers can exploit such a reformulation in order to find a global minimum (given enough CPU time). Furthermore, there exist reformulations obtained by DC-approaches (DC = difference of convex functions) as seen, for instance, in [19], where DC-functions are used to approximate the ℓ_0-norm, whereas an exact reformulation of the ℓ_0-norm (though mainly in the context of cardinality-constrained problems) was introduced in [15]. Finally, the paper [13] presents a formulation with the ℓ_0-term b being replaces by complementarity constraints.

The formulation from [13] is the basis of our approach and was already expanded in [16], where, in particular, we provide a complete equivalence of local and global minima between the original problem (SPO) and the corresponding optimization problem with complementarity constraints. Note that this is in contrast to a related approach introduced in [5] for cardinality-constrained programs where an equivalence between local minima does not hold. The recent paper [17] generalizes the technique from [13, 16] by introducing a whole class of reformulations of (SPO), all of which rely on the aforementioned idea of replacing the ℓ_0-term by suitable complementarity constraints. This class of reformulations was then used in order to motivate an exact penalty approach for the solution of SPO.

However, the exact penalty technique from [17] assumes that the general constraints $g(x) \in X$ contain nonnegativity constraints on the variables x. Otherwise, free variables have to be splitted into the positive and negative parts, which then increases the number of variables and destroys the shown one-to-one correspondence between (local) minima. Furthermore, this splitting might violate suitable constraint qualifications which are required to verify convergence results for the underlying exact penalty approach.

In this paper, we further elaborate on the general class of reformulations from [17] and, in particular, present a penalty approach which can be applied to the general problem (SPO) without the explicit assumption that the variables x are nonnegative. In particular, we show that our new penalty approach is exact in terms of (local/global) minima and, most importantly, also with respect to stationary points. In contrast to the exact penalty approach from [17], the objective function of the resulting penalty subproblems is (mildly) nonsmooth. Nevertheless, we will show that these subproblems can still be solved efficiently by suitable methods.

The paper is structured as follows: In Section 2, we state some background material from variational analysis, recall the details for the reformulation of problem (SPO) from the paper [17], and consider a so-called tightened nonlinear program associated to a local minimum of (SPO). The new exact penalty approach and related exactness results are given in Section 3. In particular, this includes exactness with respect to stationary points which, from a practical point of view, are particularly relevant. Section 4 presents two solution strategies which aim to find stationary points of the penalized subproblems, one is based on a projected gradient method, the other one on a proximal gradient-type approach. We then present some numerical results in Section 5 based on a variety of different classes of applications, namely portfolio optimization, dictionary learning, and adversarial attacks on neural networks.´We close with some final remarks in Section 6.

Notation: In the various parts of this paper, we address via

$$I_0(x) := \{i \,|\, x_i = 0\}$$

the set of indices for which x vanishes. Furthermore, we write $x \circ y$ for the Hadamard product of x and y, i.e. the componentwise multiplication of the two vectors. We abbreviate the canonical unit vector by $e_i \in \mathbb{R}^n$, indicating that the single 1 is in the i-th position, and additionally write $e := (1, 1, ..., 1)^T \in \mathbb{R}^n$. Since we will introduce sign constraints to our variables, we also denote with $\mathbb{R}^n_{\geq 0}$ the cone of vectors with only nonnegative entries in $\mathbb{R}^n$.

2 Mathematical Background

2.1 Variational Analysis

Here, we recall some results and definitions from the field of variational analysis that will be used in our subsequent theory.

To this end, let $X \subseteq \mathbb{R}^m$ be a nonempty and closed set. The *(Bouligand) tangent cone* at a point $x \in X$ to the set X is given by

$$T_X(x) := \{d \in \mathbb{R}^m \ : \ \exists d^k \to d, \exists t_k \searrow 0 \ : x + t_k d^k \in X\}.$$

The polar

$$N^F_X(x) := T_X(x)^\circ := \{v \in \mathbb{R}^m \ : \ v^T d \leq 0, \ \forall d \in T_X(x)\}$$

is the *Fréchet normal cone* of $x \in X$. The *limiting normal cone* by Mordukhovich is obtained from the Fréchet normal cone by passing to the outer limit, i.e.

$$N^{\lim}_X(x) := \limsup_{\bar{x} \to x} N^F_X(\bar{x}) := \{v \in \mathbb{R}^m \ : \ \exists x^k \to_X x, \ \exists v^k \to v \ : v^k \in N^F_X(x^k) \ \forall k \in \mathbb{N}\}.$$

On the other hand, the standard *normal cone (from convex analysis)* to a closed and convex set $X \subset \mathbb{R}^n$ is defined by

$$N_X(x) := \{v \in \mathbb{R}^n : v^T(y - x) \leq 0, \forall y \in X\}$$

for any $x \in X$. Note that all these cones coincide for convex sets, hence, we have $N_X(x) = N_X^F(x) = N_X^{\lim}(x)$ for all $x \in X$ with X closed, and convex. For the sake of completeness, all these cones are defined to be empty at any point $x \notin X$.

We also need a notion of differentiability for a nonsmooth function, for which different choices are possible. Here, we take the limiting subdifferential by Mordukhovich: Given a lower semicontinuous (lsc) function $f : \mathbb{R}^n \to \mathbb{R}$, we define

$$\partial f(x) := \{s \in \mathbb{R}^n \ : \ (s, -1) \in N_{epi(f)}^{\lim}((x, f(x)))\},$$

where epi(f) denotes the epigraph of f. We call $\partial f(x)$ the *limiting subdifferential* by Mordukhovich, and each element $s \in \partial f(x)$ a (limiting) subgradient. This limiting subdifferential has some useful calculus rules. Some of the basic ones, that will be used in our subsequent analysis, are summarized in the following result.

Lemma 2.1 (Calculus for the limiting subdifferential)**.** *We have the following calculus rules:*

(i) Let $f(x) = \sum_{i=1}^n f_i(x_i)$, where $f_i : \mathbb{R} \to \mathbb{R}$ is lsc. Then

$$\partial f(x) = \partial f_1(x_1) \times \partial f_2(x_2) \times \cdots \times \partial f_n(x_n).$$

(ii) Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be lsc. Then

$$\partial(f + \varphi)(x) \subset \partial f(x) + \partial\varphi(x).$$

(iii) Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and $\varphi : \mathbb{R}^n \to \mathbb{R}$ be lsc. Then

$$\partial\left(f + \varphi\right)(x) = \partial f(x) + \partial\varphi(x) = \nabla f(x) + \partial\varphi(x).$$

(iv) Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and $\varphi : \mathbb{R}^m \to \mathbb{R}^n$ be locally Lipschitz continuous. At a point x, set $y = \varphi(x)$. It holds

$$\partial(f \circ \varphi)(x) = \partial\left(\nabla f(y)^T\varphi\right)(x),$$

where the subdifferential is taken with respect to the function $\nabla f(y)^T\varphi : \mathbb{R}^n \to \mathbb{R}$.

(v) Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous and $g : \mathbb{R}^m \to \mathbb{R}^n$. Then

$$\partial(\varphi \circ g)(x) \subset g'(x)^T\partial\varphi(x).$$

Proof. Statement (i) (on separable functions) follows from [28, Prop 10.5]. The sum rule (ii) (with an inclusion) can be found in [23, Thm. 3.36] and results from the fact that, if one of the two functions is locally Lipschitz, then the corresponding singular subdifferential from that reference contains the zero vector only. The sum rule from (iii) (with equality) is taken from [23, Prop. 1.107]. The subsequent chain rules (iv) and (v) can be found in [23, Thm. 1.110] and [28, Thm. 10.6], respectively. □

As a subdifferential of particular interest for our purposes, we have the following one.

Lemma 2.2. *Consider the function $\varphi(x) := \rho\left\|x\right\|_0$ for some $\rho > 0$. Then*

$$\partial\varphi(x) = \left\{s \in \mathbb{R}^n \,\middle|\, s_i = 0 \text{ for all } i \text{ with } x_i \neq 0\right\}$$

holds for all $x \in \mathbb{R}^n$.

Proof. See for instance [18, 11]. □

We next introduce the dist function to an arbitrary set $X \subset \mathbb{R}^n$ as the minimal distance of x to this set, measured by an arbitrary norm, i.e.

$$\text{dist}_X(x) := \inf\{\|y - x\| \ : \ y \in X\}.$$

In case X is closed, the infimum is, of course, attained (albeit not necessarily unique), and we can replace the infimum by a minimum in this case. The dist function is famously known as an exact penalty function. To this end, we first clarifiy what we consider as exactness from here on.

Definition 2.3. *Let a problem* $\min_{x\in C} f(x)$ *be given and let* $\phi(x)$ *denote some merit function, which fulfills* $\phi(x) = 0$ *if* $x \in C$ *and* $\phi(x) > 0$ *otherwise. We call* $P_\alpha(x) := f(x) + \alpha\phi(x)$ *an* exact penalty function *if, for any (local) minimizer* x^* *of* f *over* C*, there is a finite value* $\alpha^* > 0$ *such that, for all* $\alpha \geq \alpha^*$*, the point* x^* *is also a (local) minimizer of* P_α*.*

Note that, in the previous definition and the subsequent statements within this subsection, the function f is not necessarily the objective function from our given optimization problem (SPO).

The following central result goes back to Clarke [7] and clearly illustrates why the dist function plays such an important role for exact penalty results. The statement is phrased in a way that includes the idea of not penalizing the entire feasible set, but rather some constraints with the help of the dist function.

Theorem 2.4. *Let the problem* $\min_{x\in C} f(x)$ *be given, where* $C \subset S$ *and* f *is Lipschitz of rank* L *on* S*. If* x^* *is a minimizer of* f *over* C*, then* x^* *minimizes the function* $P_\alpha(x) := f(x) + \alpha dist_C(x)$ *for all* $\alpha \geq L$ *over* S*. In addition, if* C *is closed, then, for* $\alpha > L$*, any other point minimizing* P_α *over* S *also lies in* C *(and therefore minimizes* f *over* C*).*

This theorem can be extended to local minima and a locally Lipschitz continuous function f simply by restricting $C \cap B_\varepsilon(x^*)$, where $B_\varepsilon(x^*)$ is a closed ball of radius ε around x^*, such that the local minimizer x^* solves $\min_{C\cap B_\varepsilon(x^*)} f(x)$. In this case x^* is truly a local minimizer of P_α since one can prove the existence of a neighborhood U of x^* such that

$$\text{dist}_{C\cap B_\varepsilon(x^*)}(x) = \text{dist}_C(x), \quad \forall x \in U,$$

cf. [31]. Conversely, for $\alpha > L$ any additional minimizer $\overline{x}$ of P_α over $S \cap U$ fulfills exactly the property $\text{dist}_{C\cap B_\varepsilon(x^*)}(\overline{x}) = 0 \iff \overline{x} \in C \cap B_\varepsilon(x^*)$.

Within the context of our sparse optimization problem (SPO) the feasible set is given by the preimage $g^{-1}(X)$. Consequently, we would have to apply the distance function result to the set $g^{-1}(X)$. The computationally more favorable choice, however, would be to switch to the dist function with respect to X at a point $g(x)$. This leads to an exact merit function precisely under the metric subregularity condition.

Definition 2.5. *We say that* metric subregularity *of rank* $\kappa > 0$ *is fulfilled for* $g(x) \in X$ *at a feasible point* x^**, if there is a neighborhood* U *of* x^* *such that*

$$dist_{g^{-1}(X)}(x) \leq \kappa \, dist_X(g(x)) \quad \forall x \in U.$$

As an immediate consequence, we obtain that, under metric subregularity, at least locally and for f being locally Lipschitz, the penalty function

$$\hat{P}(x) = f(x) + \alpha_* \kappa \text{dist}_X(g(x))$$

is exact. Consequently, given a local minimizer x^*, one necessarily has $0 \in \partial \hat{P}(x^*)$ by Fermat's rule for the limiting subdifferential. Application of Lemma 2.1(ii) then gives

$$0 \in \partial f(x^*) + \alpha^* \kappa \partial \text{dist}_X(g(x^*)). \tag{1}$$

The subdifferential of the dist function is well-known.

Lemma 2.6. *Let $X \subset \mathbb{R}^n$ be nonempty and closed, and let $P_X(x) := \{y \in X : \|x - y\| = dist_X(x)\}$. Then the subdifferential of the dist function ís given by*

$$\partial dist_X(x) = \begin{cases} N_X^{\lim}(x) \cap B_1(0), & x \in X \\ \frac{x - P_X(x)}{dist_X(x)}, & x \notin X. \end{cases}$$

Proof. See [28, Ex. 8.53]. □

With the necessary analysis in place, we now formulate the notion of (M-)stationarity as found in literature, cf. [20]

Definition 2.7. *Consider the problem*

$$P = \min \ f(x) \quad s.t. \quad g(x) \in X,$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz, $g : \mathbb{R}^n \to \mathbb{R}^m$ a C^1 function and $X \subset \mathbb{R}^m$ a nonempty and closed set. We call x^* a stationary point *of P if $g(x^*) \in X$ and there exists $\lambda \in N_X^{\lim}(g(x^*))$ such that*

$$0 \in \partial f(x^*) + g'(x^*)^T \lambda.$$

The result of Clarke together with the sum rule from Lemma 2.1 (ii) implies that a local minizer x^* of P satisfies

$$0 \in \partial f(x^*) + N_{g^{-1}(X)}^{\lim}(x^*),$$

which, in general, is an impractical criterion to check. Under metric subregularity, however, we arrive at the necessary optimality condition (1), where Lemma 2.6 together with Lemma 2.1 (v) infers

$$0 \in \partial f(x^*) + g'(x^*)^T \lambda, \quad \lambda \in N_X^{\lim}(g(x^*))$$

at a local minimizer x^*. Metric subregurality is therefore not only a criterion for the existence of an exact penalty function, but can also be considered a constraint qualification for a more general class of constraints.

2.2 Reformulation of Sparse Optimization Problems

Here, we recall the class of reformulations of problem (SPO) from our paper [17]. To this end, $p^\rho \colon \mathbb{R}^n \to \mathbb{R}$ be a function (usually depending on a parameter $\rho > 0$) given by

$$p^\rho(y) = \sum_{i=1}^{n} p_i^\rho(y_i) \tag{2}$$

with each $p_i^\rho \colon \mathbb{R} \to \mathbb{R}$ being such that it satisfies the following conditions:

(P.1) p_i^ρ is convex and attains a unique minimum (possibly depending on ρ) at some point $s_i^\rho > 0$;

(P.2) $p_i^\rho(0) - p_i^\rho(s_i^\rho) = \rho$;

(P.3) p_i^ρ is continuously differentiable.

Assumption (P.1) simply states that p_i^ρ is a convex function which attains its unique minimum in the open interval $(0, \infty)$. We denote this minimum by s_i^ρ. Furthermore, we write

$$m_i^\rho := p_i^\rho(s_i^\rho) \quad \text{and} \quad M^\rho := \sum_{i=1}^n m_i^\rho \tag{3}$$

for the corresponding minimal function values of p_i^ρ and p^ρ, respectively. Condition (P.2) corresponds to a suitable scaling of the function p_i^ρ that can always be guaranteed to hold by a suitable multiplication of p_i^ρ. Finally, condition (P.3) is a smoothness condition. We stress that some of our results hold under the weaker condition that p_i^ρ is only continuous like, e.g., the subsequent statements of the equivalence between global and local minima, but the observations regarding stationary points and the practical solution of the exact penalty subproblems require the continuous differentiability of the mappings p_i^ρ.

We also recall some examples of function ρ_i^ρ satisfying properties (P.1)-(P.3), cf. [17] for more details

Example 2.8. The following functions $p_i^\rho \colon \mathbb{R} \to \mathbb{R}$ satisfy properties (P.1)-(P.3):

(a) $p_i^\rho(y_i) := \rho y_i(y_i - 2)$;

(b) $p_i^\rho(y_i) := \frac{1}{2}(y_i - \sqrt{2\rho})^2$;

(c) A suitable Huber-type modification of the shifted absolute-value function $p_i^\rho(y_i) := \rho|y_i - 1|$ can also be constructed to satisfy (P.1)-(P.3). (Note that p_i^ρ itself is nonsmooth and therefore violates (P.3).)

In the following, we assume that p^ρ is given by (2) with each term p_i^ρ satisfying conditions (P.1)-(P.3) (though only continuity is required instead of (P.3) within this subsection). We then introduce the reformulation

$$\min_{x,y} f(x) + p^\rho(y), \quad \text{s.t.} \quad g(x) \in X, \ x \circ y = 0, \ y \geq 0. \tag{SPOref}$$

of problem (SPO). The following results from [17] show that (SPOref) is indeed a reformulation of the given sparse optimization problem (SPO).

Lemma 2.9. *Let p^ρ be given by (2) with each p_i^ρ satisfying properties (P.1)–(P.3), and let M^ρ be defined by (3). Then the following statements hold:*

(a) The inequality $\rho \|x\|_0 \leq p^\rho(y) - M^\rho$ holds for any feasible point (x, y) of (SPOref).

(b) Equality $\rho \|x\|_0 = p^\rho(y) - M^\rho$ holds for a feasible point (x, y) of (SPOref) if and only if $y_i = s_i^\rho$ for all $i \in I_0(x)$.

(c) If (x^*, y^*) is a local minimum of (SPOref), we have $y_i^* = s_i^\rho$ for all $i \in I_0(x^*)$.

Part (c) of the previous result is of particular interest since it tells us that, at any local minimum (x^*, y^*), the *biactive set* $\{i : x_i^* = y_i^* = 0\}$ is empty, and that the vector y^* corresponding to x^* is necessarily given by

$$y_i^* = \begin{cases} s_i^\rho, & \text{for } i \in I_0(x^*), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

We now state the essential relation between (SPO) and (SPOref) also found, see, once again, [17] for the corresponding proof.

Theorem 2.10 (Equivalence of Global Minima)**.** *A feasible x^* for (SPO) is a global minimum of (SPO) if and only if (x^*, y^*) with y^* given by (4) is a global minimum of (SPOref).*

The corresponding results for local minima also holds (recall that this is not true, in general, for a similar reformulation of cardinality-constrained problems, cf. [5]).

Theorem 2.11 (Equivalence of Local Minima)**.** *A feasible x^* for (SPO) is a local minimum of (SPO) if and only if (x^*, y^*) with y^* given by (4) is a local minimum of (SPOref).*

The previous two results imply that the sparse optimization problem (SPO) and its reformulation (SPOref) are completely equivalent programs (at least in terms of local and global minima). This motivates to solve the given nonsmooth (even discontinuous) minimization problem (SPO) by solving the equivalent smooth (continuously differentiable) program (SPOref). Of course, the latter has its difficulties, too, which arises from the complementarity constraints. Thus, its solution requires suitable problem-tailored techniques. On the other hand, we already stress at this point that the biactive set of the reformulated program is empty, which helps a lot to prove suitable properties.

2.3 The Tightened Nonlinear Program

The objectve function of (SPO) is, due to the ℓ_0-norm, not locally Lipschitz, which takes away access from standard NLP theory. To remedy this effect, we first consider the *tightened nonlinear program*

$$\min f(x) \quad \text{s.t.} \quad g(x) \in X,\ x_i = 0\ \forall i \in I_0(x^*) \tag{TNLP(x^*)}$$

around a feasible point x^*. This tightened problem plays a fundamental role for the theoretical investigation of sparse optimization problems (not from a practical point of view since the given x^* is usually a minimum and therefore unknown). For example, most (problem-tailored) constraint qualifications for the sparse nonlinear program (SPO) are typically defined by the corresponding (standard) constraint qualification for (TNLP(x^*)), cf. our previous works [16, 17] for some examples. Here, we will exploit this tightened program in a similar context.

First of all observe that x^* is a local minimum of (SPO) if and only if x^* solves the corresponding TNLP(x^*). This, in particular, motivates to define stationary points also via the tightened program.

Definition 2.12. *Let x^* be feasible for (SPO). We say that x^* is* stationary *for (SPO) if x^* is stationary for TNLP(x^*).*

If f is differentiable, this stationarity conditoins collapses to

$$0 = \nabla f(x) + \sum_{I_0(x^*)} \gamma_i e_i + g'(x)^T \lambda, \quad \gamma \in \mathbb{R}^{|I_0(x^*)|},\ \lambda \in N_X^{\lim}(g(x)).$$

In turn, at $x = x^*$, this is equivalent to

$$0 \in \nabla f(x^*) + \partial\rho\|x^*\|_0 + g'(x^*)^T\lambda, \quad \lambda \in N_X^{\lim}(g(x^*)).$$

This is *precisely* the (M-)stationarity extended to the case of (SPO). Any known result from standard NLP theory, like formulating suitable constraint qualifications and also the existence of exact penalty functions, can now be extended to (SPO) by means of the corresponding TNLP(x^*) at the point x^*.

Let us introduce that *filter matrix*

$$P \in \mathbb{R}^{|I_0(x^*)|\times n} \text{ whose columns are given by } e_i^T \ (i \in I_0(x^*)) \tag{5}$$

at a feasible point x^*. Note that P depends on the given point x^*, though this is not made precise by our notation since the underlying vector x^* will always be clear from the context. Using this matrix P, the feasible set of (TNLP(x^*)) can be rewritten as

$$C := \{x \ : \ Px = 0, \ g(x) \in X\},$$

which clarifies that P is a kind of (sparse) filter of x^*. Further writing

$$S := g^{-1}(X),$$

the feasible set C gets the representation

$$C = \{x \ : \ Px = 0, x \in S\},$$

so that C is an intersection of a linear constraint and an abstract (geometric) constraint. Theorem 2.4 now tells us that the dist function is already exact. However, we want to keep S as a constraint and penalize only the sparsity inducing part $Px = 0$. The question, whether this leads to an exact penalty formulation, is answered by the well kown error bound property from [14], specifically formulated for the (TNLP(x^*)).

Definition 2.13. *We say that the* error bound property *for* (TNLP(x^*)) *holds at* $x^* \in C$ *if there exist* $\varepsilon > 0$ *and* $\mu > 0$ *satisfying*

$$dist_C(x) \leq \mu\|Px\|$$

for all $x \in g^{-1}(X) \cap B_\varepsilon(x^*)$*, where* P *denotes the matrix from* (5)*.*

One can easily check that metric subregularity for the specific structure of C already implies the error bound property, so that we will be content with the usual constraint qualifications formulated for C which imply metric subregularity. Consequently, any conditions stronger than the error bound property yields that the penalty function

$$f(x) + \alpha \|Px\|_1, \tag{6}$$

is exact for (TNLP(x^*)). So far, extensive research has been put into constraint qualifications that imply error boundedness or are equivalent to the existence of error bounds. For example, [22] shows that a quasi-normality conditions is sufficient for an error bound. The very recent report [1] generalizes this result and shows that a slightly weaker version of quasi-normality, called directional quasi-normality, is actually equivalent to an error bound condition. Note, however, that the results are presented for functional constraints only.

Here, we view the feasibility condition $g(x) \in X$ as an abstract constraint and only penalize the seemingly difficult constraints. This will be done based on a generalized quasi-normality conditions which, in particular, holds under the more standard generalized MFCQ assumption to be introduced in the following section.

3 Exact Penalty Approach for the Sparse Optimization Problem

We recall from Theorems 2.10 and 2.11 that the given sparse optimization problem (SPO) is completely equivalent to the reformulated program (SPOref) both in terms of global and local minima. In this section, we exploit this equivalence and consider an exact penalty approach based on the structure of the reformulated program (SPOref). More precisely, we consider the penalty problem

$$\min_{x,y} f(x) + p^{\rho}(y) + \alpha |x|^T y, \quad \text{s.t.} \quad g(x) \in X, \ y \geq 0, \tag{Pen(α)}$$

where

$$|x| := (|x_1|, |x_2|, ..., |x_n|),$$

i.e. only penalize the complementarity constraints $x \circ y$ from the reformulated program (SPOref), whereas we keep the (seemingly simple) remaining constraints. Note that, and in contrast to the exact penalty approach from our previous paper [17], we do not necessarily have that the components of the vector x are nonnegative, and this is the reason why we need to take the absolute value of x within the penalized objective function. This, of course, makes this objective function nonsmooth.

To derive stationary conditions for the penalized program (Pen(α)), we first present a preliminary result which computes the subdifferential of the nonsmooth term $q(x, y) = |x|^T y$.

Lemma 3.1. *At any point (x, y), where $y \geq 0$, the subdifferential of $q(x, y) := |x|^T y$ is given by*

$$v \in \partial q(x,y) \quad \Longleftrightarrow \quad v = \begin{pmatrix} y \circ s \\ |x| \end{pmatrix}, \quad s_i \in \begin{cases} \{1\}, & x_i > 0, \\ [-1, 1], & x_i = 0, \\ \{-1\}, & x_i < 0, \end{cases} \quad \forall i = 1, ..., n.$$

Proof. First observe that the function q is continuously differentiable with respect to y with the corresponding partial derivative given by $|x|$, which then is equal to the unique second block component of the subdifferential of q. Regarding the subdifferential with respect to the variable x, we first note that the function q is separable and can be rewritten as

$$q(x, y) = \sum_{i=1}^{n} |x_i| y_i.$$

Application of the subdifferential properties from Lemma 2.1 and noting that the limiting subdifferential of a convex function coincides with the standard subdifferential from convex analysis for a convex function, we obtain the desired statement □

By application of the sum rule from Lemma 2.1 (*iii*), the limiting subdifferential of the objective function from the penalty problem (Pen(α)) simply reads

$$\partial\big(f(x) + p^{\rho}(y) + \alpha |x|^T y\big) = \begin{pmatrix} \nabla f(x) + \alpha y \circ \partial(|x|) \\ \nabla p^{\rho}(y) + \alpha |x|, \end{pmatrix}$$

and is locally bounded by continuity of ∇f and ∇p^{ρ} and boundedness of $\partial(|x|)$, which implies local Lipschitz continuity. Hence, any local minimizer (x, y) necessarily satisfies

$$0 \in \begin{pmatrix} \nabla f(x) + \alpha y \circ \partial(|x|) \\ \nabla p^{\rho}(y) + \alpha |x| \end{pmatrix} + N^{\lim}_{g^{-1}(X)}(x) + N_{\geq 0}(y)$$

where $N_{\geq 0}(y) = \{\lambda \;:\; \lambda^T(z-y) \leq 0,\ \forall z \geq 0\}$ denotes the convex normal cone to $\mathbb{R}_{\geq 0}$. From standard analysis, one has

$$\gamma \in N_{\geq 0}(y) \quad \Longleftrightarrow \quad \gamma_i \leq 0,\ \gamma_i y_i = 0.$$

With this in mind, simply writing down that stationarity condition, especially for (Pen(α)), leads to the following definition.

Definition 3.2. *We call* (x^*, y^*) stationary *for* (Pen(α)) *if* (x^*, y^*) *is feasible and there are multipliers* $\lambda \in N_X^{\lim}(g(x^*))$ *and* $\gamma_i \geq 0$ $\big(i \in I_0(y^*)\big)$ *with*

$$0 \in \begin{pmatrix} \nabla f(x^*) + \alpha y^* \circ \partial(|x^*|) + g'(x^*)^T\lambda \\ \nabla p^\rho(y^*) + \alpha|x^*| - \sum_{i\in I_0(y^*)} \gamma_i e_i \end{pmatrix}. \tag{7}$$

The following result essentially shows that (6) is an exact penalty function of the tightend program (TNLP(x^*)) if and only if (Pen(α)) is an exact penalty function of reformulated program (SPOref).

Theorem 3.3. *Let* x^* *be a local minimum of* (SPO) *or, equivalently,* (x^*, y^*) *be a local minimum of* (SPOref) *with* y^* *being given by* (4)*. Then* (6) *is an exact penalty function of* (TNLP(x^*)) *at* x^* *if and only if* (Pen(α)) *is an exact penalty function of* (SPOref) *at* (x^*, y^*)*.*

Proof. Recall that Lemma 2.9 implies $y_i^* = s_i^\rho$ for $i \in I_0(x^*)$ and $y_i^* = 0$ otherwise. In particular, we therefore have $\{1, ..., n\} = I_0(x^*) \cup I_0(y^*)$.

We first assume that (6) be exact for TNLP(x^*). Then we can find a neighborhood $U = U_x \times U_y$ around (x^*, y^*) and constants σ_x, $\sigma_y > 0$ such that implication

$$(x, y) \in U \quad \Rightarrow \quad |x_i| > \sigma_x,\ y_j > \sigma_y\ \forall (i,j) \in I_0(y^*) \times I_0(x^*),$$

holds and, additionally, x^* is a local minimizer of (6) in U_x for all sufficiently large α. Now take an arbitrary element $(x, y) \in U$. Choose $\overline{y}$ such that $\overline{y}_i = y_i$ for $i \in I_0(x^*)$ and $\overline{y}_i = 0$ for $i \in I_0(y^*)$. It follows that

$$\|(x, \bar{y}) - (x^*, y^*)\| \leq \|(x, y) - (x^*, y^*)\|\,,$$

so that we also have $(x, \overline{y}) \in U$. Furthermore, it follows that

$$\begin{aligned} p^\rho(\bar{y}) = \sum_{i=1}^n p_i^\rho(\bar{y}_i) \; &= \sum_{i\in I_0(x^*)} p_i^\rho(\bar{y}_i) + \sum_{i\in I_0(y^*)} p_i^\rho(\bar{y}_i) \; = \sum_{i\in I_0(x^*)} p_i^\rho(y_i) + \sum_{i\in I_0(y^*)} p_i^\rho(0) \\ &\geq \sum_{i\in I_0(x^*)} p_i^\rho(s_i^\rho) + \sum_{i\in I_0(y^*)} p_i^\rho(0) \; = \sum_{i\in I_0(x^*)} p_i^\rho(y_i^*) + \sum_{i\in I_0(y^*)} p_i^\rho(y_i^*) = p^\rho(y^*), \end{aligned} \tag{8}$$

where the inequality and the subsequent equality follow from the definitions of the minima s_i^ρ and the elements y_i^*, respectively. Moreover, we can measure a deviation of y to $\overline{y}$ with the help of the penalty term p^ρ. To this end, fix an index $i \in I_0(y^*)$. Then

$$p_i^\rho(y_i) - p_i^\rho(\overline{y}_i) + \alpha|x_i|y_i - \alpha|x_i|\overline{y}_i = (\nabla p_i^\rho(\xi_i) + \alpha|x_i|)\, y_i \geq (\nabla p_i^\rho(0) + \alpha\sigma_x)\, y_i, \tag{9}$$

where the equation follows from the differential mean-value theorem together with $\bar{y}_i = 0$ for $i \in I_0(y^*)$, and the inequality results from the choice of σ_x in combination with the fact that

$\nabla p_i^\rho(\xi_i) \geq \nabla p_i^\rho(0)$ due to the convexity of each p_i^ρ since this implies the monotonicity of its derivate. We can assure the right-hand side of the previous expression to be nonnegative by taking

$$\alpha \geq \frac{-\nabla p_i^\rho(0)}{\sigma_x} \quad \forall i \in I_0(y^*)$$

(note that the lower bounds are positive numbers since the functions p_i^ρ are assumed to attain their unique minimum within the interval $(0, +\infty)$). This choice of α then guarantees that

$$p_i^\rho(y_i) - p_i^\rho(\overline{y}_i) + \alpha|x_i|y_i - \alpha|x_i|\overline{y}_i \geq 0 \quad \forall i = 1, \dots, n \tag{10}$$

as this follows from (9) for all $i \in I_0(y^*)$ (recall that we have $y_i \geq 0$), whereas this holds obviously for all $i \in I_0(x^*)$ since $\bar{y}_i = y_i$ for these indices. Now let $\overline{\alpha}$ be the very outer penalty parameter that correlates to the exactness of (6). Choose

$$\alpha^* \geq \max\left\{\frac{\overline{\alpha}}{\sigma_y}, \max_{i \in I_0(y^*)} \frac{-\nabla p_i^\rho(0)}{\sigma_x}\right\}.$$

For any $\alpha \geq \alpha^*$, we then obtain

$$f(x) + p^\rho(y) + \alpha|x|^T y \geq f(x) + p^\rho(\overline{y}) + \alpha|x|^T\overline{y} \geq f(x) + p^\rho(y^*) + \alpha\sigma_y\|Px\|_1 \geq f(x^*) + p^\rho(y^*),$$

where the first inequality exploits (10), the second one uses (8) together with

$$|x|^T\overline{y} = \sum_{i=1}^n |x_i|\overline{y}_i = \sum_{i \in I(x^*)} |x_i|\overline{y}_i = \sum_{i \in I(x^*)} |x_i|y_i \geq \sigma_y \sum_{i \in I(x^*)} |x_i| = \sigma_y\|Px\|_1,$$

and the final estimate takes into account the assumed exactness of (6) at x^* (note that $Px^* = 0$ by definition of the matrix P). This verifies the exactness of (Pen(α)).

Conversely, assume that (Pen(α)) is exact for (SPOref) at a local minimizer (x^*, y^*) with neighborhood U, again in the form $U = U_x \times U_y$. Hence, we have

$$f(x) + p^\rho(x) + \alpha|x|^T y \geq f(x^*) + p^\rho(x^*) + \alpha|x^*|^T y^* = f(x^*) + p^\rho(x^*) \tag{11}$$

for all $(x, y) \in U_x \times U_y$ satisfying $g(x) \in X$ and $y \geq 0$, where the equation follows from the fact that $\alpha|x^*|^T y^*$ by definition of y^*. Now, observe that

$$f(x) + \alpha\|Px\|_1 = f(x) + \alpha \sum_{i \in I_0(x^*)} |x_i| \geq f(x) + \frac{\alpha}{\max_{j \in I(x^*)} y_j^*} \sum_{i \in I_0(x^*)} y_i^*|x_i|. \tag{12}$$

Let $\overline{\alpha}$ denote the outer penalty parameter that correlates to the exactness of (Pen(α)). Set

$$\alpha^* = \overline{\alpha} \cdot \max_{j \in I_0(x^*)} y_j^*$$

and take $\alpha \geq \alpha^*$. Choose and arbitrary $x \in U_x$ satisfying $g(x) \in X$. It follows that $(x, y^*) \in U$. Hence, we can apply (11) to this vector pair and obtain

$$f(x) + p^\rho(y^*) + \alpha|x|^T y^* \geq f(x^*)p^\rho(y^*),$$

which simplifies to

$$f(x) + \alpha|x|^T y^* \geq f(x^*). \tag{13}$$

Altogether, we therefore have

$$
\begin{aligned}
f(x) + \alpha \|Px\|_1 &\geq f(x) + \frac{\alpha}{\max_{j \in I(x^*)} y_j^*} \sum_{i \in I_0(x^*)} y_i^* |x_i| \\
&= f(x) + \frac{\alpha}{\max_{j \in I(x^*)} y_j^*} |x|^T y^* \\
&\geq f(x) + \frac{\alpha^*}{\max_{j \in I(x^*)} y_j^*} |x|^T y^* \\
&= f(x) + \overline{\alpha} |x|^T y^* \\
&\geq f(x^*).
\end{aligned}
$$

where the first inequality is taken from (12), the subsequent equation follows from the definition of y^*, the next estimate exploits that $\alpha \geq \alpha^*$, afterwards we use the definition of α^*, and the final inequality results from the exact penalty property in (13). This proves the exactness of (6). □

Exactness of (Pen(α)) is therefore equivalent to the exactness of (6). Consequently, it suffices to consider constraint qualifications for (TNLP(x^*)) that imply the desired error bounds. As mentioned before, we will present the following two conditions.

Definition 3.4. *Let x^* be feasible for* (SPO). *We say that the*

(i) sparse quasi-normality *is satisfied at x^*, if there is no nonzero $\lambda = (\lambda^a, \lambda^b)$ such that*

$$P^T \lambda^a + g'(x^*)^T \lambda^b = 0$$

holds and that satisfies the following condition: there is $\{x^k\} \to x^*$, $\{y^k\} \to g(x^*)$ and $\{\lambda^k\} \to \lambda^b$ such that

$$\lambda^k \in N_X^F(y^k), \quad \lambda_i^a \neq 0 \;\Rightarrow\; \lambda_i^a x_i^k > 0, \quad \lambda_i^b \neq 0 \;\Rightarrow\; \lambda_i^b (g_i(x^k) - y_i^k) > 0.$$

(ii) sparse generalized Mangasarian Fromovitz *(SP-GMFCQ) is satisfied at x^*, if there is no nonzero $\lambda = (\lambda^a, \lambda^b)$ such that*

$$P^T \lambda^a + g'(x^*)^T \lambda^b = 0, \quad (\lambda^a, \lambda^b) \in \mathbb{R}^{|I_0(x^*)|} \times N_X^{lim}(x^*).$$

Obviously, quasi-normality implies the corresponding MFCQ condition, from which we can recover the desired error bounds as well as constraint qualifications to obtain stationarity at a local minimizer.

Theorem 3.5. *Let x^* be feasible for* (SPO) *and assume sparse quasi-normality or SP-GMFCQ to hold at x^*. Then the metric subregularity for*

$$(Px, g(x)) \in \{0\}^{|I_0(x^*)|} \times X$$

is satisfied at x^*. Furthermore, the error bound property for (TNLP(x^*)) *is satisfied at x^*.*

Proof. Compare again [2]. □

In the common setting, where one has $X = \mathbb{R}^m_{\leq 0}$, the SP-GMFCQ collapses to SP-MFCQ as presented in our previous work.

Exactness results, as mentioned before, are typically formulated in terms of local minima. From a practical point of view, however, corresponding results on stationary points are significantly more important, though much less investigated. The following result shows that we also have exactness in terms of stationary points.

Theorem 3.6. *Let (x^*, y^*) be stationary for (SPOref). Then there exists an $\alpha^* > 0$ such that (x^*, y^*) is a stationary point of (Pen(α)) for all $\alpha \geq \alpha^*$.*

Proof. By stationarity of (x^*, y^*) for (SPOref), we have

$$0 = \nabla f(x^*) + g'(x^*)^T \lambda^* + \sum_{i \in I_0(x^*)} \gamma_i^x y_i^* e_i,$$
$$0 = \nabla p^\rho(y^*) + \sum_{i \in I_0(y^*)} (\gamma_i^y x_i^* - \nu_i^*) \, e_i$$

for suitable Lagrange multipliers $\lambda^*, \gamma^x, \gamma^y, \nu^*$, where $\lambda^* \in N_X(g(x^*))$ and $\nu_i^* \geq 0$. Let

$$\alpha \geq \alpha^* := \max \left\{ \max_{i \in I_0(x^*)} |\gamma_i^x|, \max_{i \in I_0(y^*)} |\gamma_i^y| \right\}.$$

Then, clearly, we have

$$\gamma_i^x y_i^* \in \alpha y_i^* \partial |x_i^*| \quad \forall i \in I_0(x^*)$$

since $\partial |x_i^*| = [-1, 1]$. Furthermore. for each $i \in I_0(y^*)$, we get

$$0 = \nabla p_i^\rho(0) + \gamma_i^y x_i^* - \nu_i^* = \nabla p_i^\rho(0) + \alpha |x_i^*| - (\nu_i^* + \underbrace{\alpha |x_i^*| - \gamma_i^y x_i^*}_{\geq 0}).$$

Now set $\lambda = \lambda^*$, $\gamma_i = \nu_i^* + \alpha |x_i^*| - \gamma_i^y x_i^* \geq \nu_i^* \geq 0$ for $i \in I_0(y^*)$ and $\gamma_i = 0$ otherwise. The tuple $(x^*, y^*, \lambda, \gamma)$ satisfies (7) and (x^*, y^*) is stationary for (Pen(α)). □

The following result contains an exactness statement for the other direction. This result may be viewed as a generalization of a related theorem given in [27] in the context of mathematical programs with equilibrium constraints (MPECs), though our assumptions are weaker.

Theorem 3.7. *Let (x^*, y^*) be a stationary point of (SPOref) such that SP-GMFCQ holds at x^*. Then there exists an $\alpha^* > 0$ and a neighborhood U of (x^*, y^*) such that for all $\alpha \geq \alpha^*$, every stationary point of (Pen(α)) in U is a stationary point of (SPOref).*

Proof. Assume, by contradiction, that there is a sequence $\alpha_k \to \infty$ and a sequence $\{(x^k, y^k)\}$ such that $(x^k, y^k) \to (x^*, y^*)$, where (x^k, y^k) is stationary for (Pen(α)) with $\alpha = \alpha_k$, but not stationary for (SPOref). The proof will be carried out in two parts.

(i) We show that y_i^k terminates in 0 for sufficiently large k if $i \in I_0(y^*)$.

(ii) We show that, under SP-MFCQ, x_i^k also terminates in 0 for sufficiently large k if $i \in I_0(x^*)$.

The strict complementarity between x^k and y^k is then a consequence of strict complementarity of (x^*, y^*). It follows that (x^k, y^k) is feasible for (SPOref) and, by stationarity with respect to (Pen(α)), therefore also stationary for (SPOref), which then finishes the contradiction.

(i): Let $i \in I_0(y^*)$ and assume $y_i^k \searrow 0$. By stationarity of (Pen(α)) we have

$$0 = \nabla p_i^\rho(y_i^k) + \alpha_k |x_i^k|.$$

Since (x^*, y^*) is stationary for (SPOref), we have, by strict complementarity, that $x_i^* \neq 0$ so that $\alpha_k |x_i^k| \to \infty$, whereas $\nabla p_i^\rho(y_i^k)$ remains bounded and the right hand side in the above equation diverges, which is a contradiction. This implies that $y_i^k = 0$ for almost all k, if $y_i^* = 0$.

(ii): We also claim that $x_i^k = 0$ for all $i \in I_0(x^*)$ and all k sufficiently large. To this end, note that tationarity of (x^k, y^k) for (Pen(α)) can be written as

$$0 = \nabla f(x^k) + g'(x^k)^T \lambda^k + \alpha_k \sum_{i \in I_0(x^*)} \gamma_i^k e_i + \alpha_k \sum_{i \in I_0(y^*)} \gamma_i^k e_i, \tag{14}$$

where $\gamma_i^k \in y_i^k \partial(|x_i^k|)$ and $\lambda^k \in N_X^{\lim}(g(x^k))$. By part (i), we immediately have $\gamma_i^k = 0$ for $i \in I_0(y^*)$ and k sufficiently large. Assume now there is $i \in I_0(x^*)$ such that $|x_i^k| \searrow 0$. Then, in particular, $\gamma_i^k \in \{-y_i^k, y_i^k\}$, where $y_i^k \to y_i^* \neq 0$. In particular, this implies that the sequence $\{\|(\lambda^k, \alpha_k \gamma^k)\|\}$ converges to ∞ since $\alpha_k \to \infty$. Without loss of generality, we may assume that the corresponding normalized sequence converges, say

$$(\bar{\lambda},\ \bar{\gamma}) = \lim_{k \to \infty} \frac{1}{\|(\lambda^k,\ \alpha_k \gamma^k)\|} (\lambda^k,\ \alpha_k \gamma^k).$$

Using the cone property of $N_X^{\lim}$ as well as the robustness of the limiting normal cone, see [24, Prop. 1.3], we have $\bar{\lambda} \in N_X^{\lim}(g(x^*))$. Dividing (14) with $\|(\lambda^k,\ \alpha_k \gamma^k)\|$ and pushing to the limit $k \to \infty$, we therefore get

$$0 = g'(x^*)^T \bar{\lambda} + P^T \bar{\gamma}, \quad (\bar{\lambda},\ \bar{\gamma}) \neq 0,$$

a contradiction to the assumed SP-GMFCQ. Hence, we have $x_i^k = 0$ for all $i \in I_0(x^*)$ and all sufficiently large k.

As a consequence of parts (i) and (ii), it follows that, eventually, $I_0(x^k) = I_0(x^*)$ and $I_0(y^k) = I_0(y^*)$ holds so that, in fact, (x^k, y^k) is strictly complementary and, in particular, feasible for (SPOref).

It remains to see that y^k takes the values $y_i^k = s_i^\rho$ $(= y_i^*)$ for $i \notin I_0(y^*)$. This is true since, for all $i \notin I_0(y^*) = I_0(y^k)$, we obtain from (7) that

$$0 = \nabla p_i^\rho(y_i^k)$$

holds, so that y_i^k is a minimum of the convex function p_i^ρ. However, by assumption, p_i^ρ attains its unique minimum at s_i^ρ, so that $y_i^k = s_i^\rho$ follows for all sufficiently large k. This finishes the proof. □

The final result of this section is not directly connected to the previous exactness results. In a numerical setting, increasing the penalty parameter α should, regardless of exactness (or the satisfaction of suitable constraint qualifications), minimize the residual $|x|^T y$. This is, in fact, the case even if we are only able to find approximate stationary points.

Theorem 3.8. *Let $\varepsilon > 0$ such that there exists a $t^* > 0$ for which $\nabla p_i^\rho(t^*) > \varepsilon$. Let $\{(x^k, y^k)\}$, $\{z^k\}$ be sequences satisfying*

$$g(x^k) + z^k \in X, \quad z^k \in B_\varepsilon(0), \quad y_i^k, \geq 0$$

and assume that there are sequences $\{\lambda^k\} \subset N_X^{\lim}(g(x^k) + z^k)$ and $\gamma^k \in N_{\geq 0}(y^k)$ such that

$$\begin{pmatrix} -\nabla f(x^k) - \alpha_k y^k \circ \partial(|x^k|) - g(x^k)^T \lambda^k \\ -\nabla p^\rho(y^k) - \alpha_k |x^k| + \sum_{i \in I_0(y^k)} \gamma_i^k e_i \end{pmatrix} \in B_\varepsilon(0)^2.$$

If $\alpha_k \to \infty$, then necessarily $|x^k| \circ y^k \to 0$.

Proof. It suffices to consider the assumption

$$-\nabla p^\rho(y^k) - \alpha_k |x^k| + \sum_{i \in I_0(y^k)} \gamma_i^k e_i \in B_\varepsilon(0).$$

This implies the existence of another bounded sequence $\{\delta^k\} \subset B_\varepsilon(0)$ such that

$$0 = \nabla p^\rho(y^k) + \alpha_k |x^k| - \sum_{i \in I_0(y^k)} \gamma_i^k e_i + \delta^k.$$

Now multiply the above equation componentwise with y^k. We then obtain for each $i \notin I_0(y^k)$ (otherwise there is nothing to show)

$$0 = y_i^k \nabla p_i^\rho(y_i^k) + \alpha_k y_i^k |x_i^k| + y_i^k \delta_i^k > y_i^k(\nabla p_i^\rho(y_i^k) - \varepsilon) + \alpha_k y_i^k |x_i^k|.$$

Assume $y_i^k \to \infty$. Then $y_i^k > t^*$ and, by assumption and convexity of p_i^ρ, we then have $\nabla p_i^\rho(y_i^k) - \varepsilon > 0$, which leads to a contradiction. Hence y_i^k is bounded. Now assume there is a subsequence K for which $y_i^k |x_i^k| > c$. Then $\alpha_k y_i^k |x_i^k| \to_K \infty$, which again is a contradiction. Altogether, this shows that $|x^k|^T y^k \to 0$ for $k \to \infty$. □

We only require the iterates y^k from the exact penalty method to be nonnegative and $\nabla p_i^\rho(t)$ to become sufficiently large. Both is very easy to satisfy, since, on the on hand, we can simply project any iterate y^k onto $\mathbb{R}^n_{\geq 0}$ and, on the other hand, choose ∇p^ρ in a way such that the derivative assumption from the previous result holds. Note that this assumption is satisfied for all mappings from Example 2.8.

4 Solution Strategies for Penalized Problems

The solution of the penalized subproblems requires a beneficial structure of the constraints $g(x) \in X$. For example, for $X = \mathbb{R}^n_{\leq 0}$, one could simply augment the constraints by changing the objective function to

$$F(x) = f(x) + \mu \max\{0, g(x)\}^2,$$

and incorporate an augmented Lagrangian method. On the other hand, for $g(x) = x$ and the set X being nonempty, closed and convex, which admits a suitable projection, one could approach (Pen(α)) with a projected gradient or proximal gradient method. We are mostly concerned with the latter scenario. To this end, we first consider an unconstrained version of (Pen(α)) and present two unique ways of applying first-order methods to the penalty formulation. These methods have to be slightly adapted in a practical setting to accommodate for a possible projection, which was no issue in our numerical test examples.

4.1 A Projected Spectral Gradient Method

We rewrite problem (Pen(α)) with the introduction of an auxiliary variable s in the form

$$\min_{x,s,y} f(x) + p^\rho(y) + \alpha s^T y, \quad \text{s.t.} \quad y \geq 0,\ |x| \leq s. \tag{15}$$

The problem (15) has a convex feasible set and is equivalent to (Pen(α)) in the sense that both objective functions take the same exact value if, for $y_i \neq 0$, we have $|x_i| = s_i$, which is also necessary at a local minimum of (15). However, $|x_i| < s_i$ does not change the target value of (15) whenever $y_i = 0$, which shows that (15), in general, produces infinitely many local minima, so one may be interested in using the more rigorous constraint $|x| = s$. This feasible set, however, is no longer convex and computationally less preferable. We rather propose to simply modify a solution approach to (15) in such a way that all iterates satsify the constraint $|x| = s$, which is not hard to achieve since simply overwriting s to the value of $|x|$ leads, in any case, to a better target value of (15).

As mentioned before, the set $\Omega := \{(x, s, y) \ : \ |x| \leq s,\ y \geq 0\}$ is convex since, by reordering the entries of (x, s, y), one may regard Ω as the Cartesian product

$$\Omega = \Pi_{i=1}^n \text{epi}(|\cdot|) \times \mathbb{R}^n_+,$$

where epi($|\cdot|$) denotes the epigraph of the absolute value function:

$$\text{epi}(|\cdot|) := \{(x, s) \ : \ |x| \leq s\}.$$

The latter is convex since the absolute value function is convex. Then it is easy to compute the projection onto Ω.

Lemma 4.1. *Let $\Omega = \{(x, s, y) \ : \ |x| \leq s,\ y \geq 0\}$. Then the projection $(x, s, y) = P_\Omega(u, v, w)$ is given by*

$$(x_i, s_i) = \begin{cases} (u_i, v_i), & \text{if} \quad |u_i| \leq v_i, \\ (0, 0), & \text{if} \quad |u_i| \leq -v_i, \\ \frac{1}{2}(u_i + v_i, u_i + v_i), & \text{if} \quad u_i > |v_i|, \\ \frac{1}{2}(u_i - v_i, v_i - u_i), & \text{if} \quad u_i < -|v_i|, \end{cases}$$

and $y = \max\{w, 0\}$.

The following algorithm is mainly the spectral gradient method (SPG) as presented in [4], which we copy over for the most part, but overwrite the values of s^k, obtained in each iteration, by the preferable value of $|x^k|$. For ease of notation, we will refer to the target function of (15) by $F(x, s, y)$.

Algorithm 4.2 (SPG)**.** *Let $\beta \in (0, 1)$, $0 < \sigma_{\min} < \sigma_{\max}$, and M be a positive integer. Let $(x^0, s^0, y^0) \in \Omega$ be an arbitrary initial point. For $k = 0, 1, 2, ...$*

1. *Compute $\sigma_k^{SPG} \in [\sigma_{\min}, \sigma_{\max}]$ by*

$$\sigma_k^{SPG} = \begin{cases} 1 & \text{if} \quad k = 0, \\ \max\left\{\sigma_{\min}, \min\left\{\frac{(v^k)^T w^k}{(v^k)^T v^k}, \sigma_{\max}\right\}\right\} & \text{otherwise,} \end{cases}$$

where $v^k = (x^k, s^k, y^k) - (x^{k-1}, s^{k-1}, y^{k-1})$ and $w^k = \nabla F(x^k, s^k, y^k) - \nabla F(x^{k-1}, s^{k-1}, y^{k-1})$, and set

$$d^k := P_\Omega\left((x^k, s^k, y^k) - \frac{1}{\sigma_k^{SPG}} \nabla F(x^k, s^k, y^k)\right) - x^k$$

2. *Set* $t \leftarrow 1$ *and* $F_k^{ref} = \max\{F(x^{k-j+1}, s^{k-j+1}, y^{k-j+1}) \ : \ 1 \leq j \leq \min\{k+1, M\}\}$. *If*

$$F((x^k, s^k, y^k) + td^k) \leq F_k^{ref} + t\beta \nabla F(x^k, s^k, y^k)^T d^k, \tag{16}$$

set $t_k = t$ *and set*

$$\begin{aligned} x^{k+1} &= x^k + t_k (d^k)_{1:n}, \\ s^{k+1} &= |x^{k+1}|, \\ y^{k+1} &= y^k + t_k (d^k)_{2n+1:3n}, \end{aligned} \tag{17}$$

and finish the iteration. Otherwise, choose $t_{new} \in [0.1t, 0.5t]$, *set* $t \leftarrow t_{new}$ *and repeat the test (16).*

Notice that the update of s^k by $|x^k|$ is in line with [4] as it is easy to see that in any case for a feasible tuple (x, s, y) it holds $F(x, |x|, y) \leq F(x, s, y)$. By continuity of the absolute value function, any accumulation point (x^*, s^*, y^*) obtained from (4.2), in particular, satisfies $s^* = |x^*|$. Notice that (15) is, in fact, a smooth NLP, as one may simply write

$$(x, s) \in \text{epi}(|\cdot|) \quad \Leftrightarrow \quad x - s \leq 0 \ \wedge \ -x - s \leq 0.$$

This implies that any other suitable smooth NLP solver can be used and, under additional constraints of the type $h(x) = 0, \ g(x) \leq 0$, Lagrangian-type methods are applicable or Lagrangian stationarity can be used to derive the Euclidean projection onto the feasible set.

4.2 A Proximal Point Method

We may also regard the problem (Pen(α)) as a composite optimization problem, where $f_1(x, y) := f(x) + p^\rho(y)$ is the smooth part and $f_2(x, y) := \alpha |x|^T y$ the nonsmooth function for which we have to compute the prox-operator

$$\text{prox}_\gamma^{SP}(u, v) := \arg\min_{\substack{x,y \\ y \geq 0}} f_2(x, y) + \frac{1}{2\gamma} \|(x, y) - (u, v)\|_2^2 .$$

By separability of f_2, we have

$$f_2(x, y) + \frac{1}{2\gamma} \|(x, y) - (u, v)\|_2^2 = \sum_{i=1}^n \alpha |x_i| y_i + \frac{1}{2\gamma} \left((x_i - u_i)^2 + (y_i - v_i)^2 \right) .$$

At a solution (x^*, y^*), we can always infer $\text{sign}(x_i^*) = \text{sign}(u_i)$ so that we only have to solve the optimization problem

$$\min_{x,y} \alpha xy + \frac{1}{2\gamma} \left((x - u)^2 + (y - v)^2 \right) \quad \text{s.t.} \quad x \geq 0, \ y \geq 0, \tag{18}$$

where $u \geq 0$.

Lemma 4.3. *Let* $u \geq 0$ *and* $v \in \mathbb{R}$. *The solution to problem (18) is given by*

$$(x, y) = \begin{cases} (u, 0), & \textit{if} \quad v < 0, \\ \frac{1}{1-\gamma^2\alpha^2}(u - \gamma\alpha v, v - \gamma\alpha u), & \textit{if} \quad \frac{1}{\gamma} > \alpha \ \textit{and} \ u \in [\gamma\alpha v, \frac{v}{\gamma\alpha}], \\ (0, v)\frac{sign(v-u)+1}{2} + (u, 0)\frac{1-sign(v-u)}{2}, & \textit{otherwise.} \end{cases}$$

Proof. Assume $v < 0$. For a fixed x, the best possible value of y for (18) is clearly $y = 0$. In that case, however, x is necessarily given by $x = u$.

Assume for the remainder $v \geq 0$. First notice that (18) is coercive on the feasible set which implies that a solution always exist. We discern the following cases.

1. Let $1/\gamma > \alpha$, then the target function of (18) is uniformly convex. Furthermore, for $u \in [\gamma\alpha v, \frac{v}{\gamma\alpha}]$, we have

$$1 - \gamma^2\alpha^2 > 0, \quad u - \gamma\alpha v \geq 0, \quad v - \gamma\alpha u \geq 0,$$

   and the point

$$(x, y) = \frac{1}{1 - \gamma^2\alpha^2}(u - \gamma\alpha v, v - \gamma\alpha u)$$

   is both a stationary and feasible point of the objective function and hence the solution.

2. Let $1/\gamma > \alpha$ and $u \notin [\gamma\alpha v, \frac{v}{\gamma\alpha}]$. Then the target function of (18) has only one (unconstrained) stationary point, which is not feasible. Therefore, the solution must lie on the boundary, so that either $(x, y) = (u, 0)$ if $v < u$ or $(x, y) = (0, v)$ if $u < v$.

3. Let $1/\gamma = \alpha$. Then the target function has infinitely many stationary points if and only if $u = v$ with condition $(x + y) = u$, or there are no (unconstrained) stationary points. In any case, the minimal value is attained at a boundary point $(x, y) = (u, 0)$ if $v < u$ or $(x, y) = (0, v)$ if $u \leq v$.

4. Finally, let $1/\gamma < \alpha$. The target function has only one stationary point, which is necessarily a saddle point. The only solutions are found on the boundary so that again $(x, y) = (u, 0)$ if $v < u$ or $(x, y) = (0, v)$ if $u \leq v$.

Altogether the desired representation follows. □

The proximal gradient method is then implemented with a nonmonotone stepsize as for instance seen in [9].

Note that we have only stated the projection and, respectively, the proximal point operator in an unconstrained case, to show that either method is very well realizable. Depending on the optimization problem, of course, modifications are necessary, or additional constraints have to, for instance, be augmented and solved via an augmented Lagrangian approach.

5 Numerical Results

5.1 Sparse Portfolio Optimization

5.1.1 Problem Formulation

To measure the quality of a portfolio of investments, one possible formulation of the Markov model [21] takes an estimated risk, given by some covariance matrix Q, and compares it to a possible return, given by some vector μ. The investor has a certain amount of capital to work with, and we assume, additionally, that short sales are possible. We therefore consider the problem

$$\min_x \ \frac{1}{2}x^TQx - \beta\mu^Tx + \rho\,\|x\|_0 \quad \text{s.t.} \quad e^Tx = 1. \tag{19}$$

Since the investor might be dissuaded from portfolio strategies with a wide spread, a sparsity-inducing penalty term is in place. Using the ℓ_1-norm, at any feasible point, we have

$$\|x\|_1 = \sum_{x_i \geq 0} x_i + \sum_{x_i < 0} |x_i|, \quad \sum_{x_i \geq 0} x_i - \sum_{x_i < 0} |x_i| = 1.$$

Hence, a spread without shorting always leads to $\|x\|_1 = 1$, with possibly no sparsity in x. For a spread with shorting, on the other hand, we necessarily have $\|x\|_1 > 1$. It follows that the ℓ_1-regularization is a bad choice, since it favors spreads without shorting and possibly with no sparsity.

It is well known that the above problem (19) can be represented as a quadratic mixed integer problem of the form

$$\min_{\substack{x \in \mathbb{R}^n, \\ y \in \{0,1\}^n}} \frac{1}{2} x^T Q x - \beta \mu^T x + \rho \|y\|_1 \quad \text{s.t.} \quad e^T x = 1, \ y_i l_i \leq x_i \leq y_i u_i, \ i = 1, ..., n,$$

where l and u denote some (sufficiently large) lower and upper bounds satisfying $l \leq 0 \leq u$. Specialized combinatorial solver can be applied to solve the mixed integer formulation, for instance, Gurobi.

To solve (19) with our exact penalty approach, we focus on the spectral gradient method. The reason for this is that we require a projection onto the set

$$X = \{(x, s) : \ e^T x = 1, \ x - s \leq 0, \ -x - s \leq 0\}. \tag{20}$$

Assume we want to project the vector (a, b). Necessary and sufficient for the solution (x^*, s^*) is the existence of multipliers $\lambda^+, \ \lambda^- \in \mathbb{R}^n_{\geq 0}$ and $\mu \in \mathbb{R}$ that fulfill

$$\begin{pmatrix} x^* - a + \mu e + \sum \lambda_i^+ e_i - \sum \lambda_i^- e_i \\ s^* - b - \sum \lambda_i^+ e_i - \sum \lambda_i^- e_i \end{pmatrix} = 0, \tag{21}$$

$$e^T x^* = 1, \tag{22}$$

$$\lambda^+ \geq 0, \quad s^* - x^* \geq 0, \quad (s^* - x^*)^T \lambda^+ = 0, \tag{23}$$

$$\lambda^- \geq 0, \quad s^* + x^* \geq 0, \quad (s^* + x^*)^T \lambda^- = 0. \tag{24}$$

A simple calculation shows that we require

$$s^* - x^* - 2\lambda^+ = b - a + \mu e, \quad \text{and} \quad s^* + x^* - 2\lambda^- = a + b - \mu e.$$

From the complementarity in (23) and (24), we have

$$\begin{aligned} s^* - x^* &= \max\{0, b - a + \mu e\}, & s^* + x^* &= \max\{0, a + b - \mu e\}, \\ 2\lambda^+ &= \max\{0, a - b - \mu e\}, & 2\lambda^- &= \max\{0, \mu e - a - b\}. \end{aligned}$$

By (22), it holds

$$e^T (s^* + x^*) - e^T (s^* - x^*) = 2$$

or, equivalently, we need μ as a root of

$$t(u) := \sum_i \max\{0, a_i + b_i - u\} - \max\{0, b_i - a_i + u\} - 2. \tag{25}$$

It is easy to see that t is continuous and monotonically decreasing. A root μ of t must exist, since clearly there is a solution to (21)-(24). Set

$$u_- := \max\{\max_i\{a_i + b_i\}, \max_i\{a_i - b_i\}\}, \quad u_+ := \min\{\min_i\{a_i + b_i\}, \min_i\{a_i - b_i\}\} - \frac{2}{n}.$$

Plugging the two values into t, one has $t(u_-) \leq 0$ and $t(u_+) \geq 0$ and, therefore, u_- and u_+ can be used as the start of a bisection method to obtain μ. We have the following Lemma.

Lemma 5.1. *Consider the set X in (20) and let $(x^*, s^*) = P_X(a, b)$ denote the euclidean projection of (a, b) onto X. Then*

$$\begin{aligned} x^* &= \frac{1}{2}\left(\max\{0, a + b - \mu e\} - \max\{0, b - a + \mu e\}\right), \\ s^* &= \frac{1}{2}\left(\max\{0, b - a + \mu e\} + \max\{0, a + b - \mu e\}\right), \end{aligned}$$

where μ is a root of t as given in (25).

5.1.2 Numerical Tests

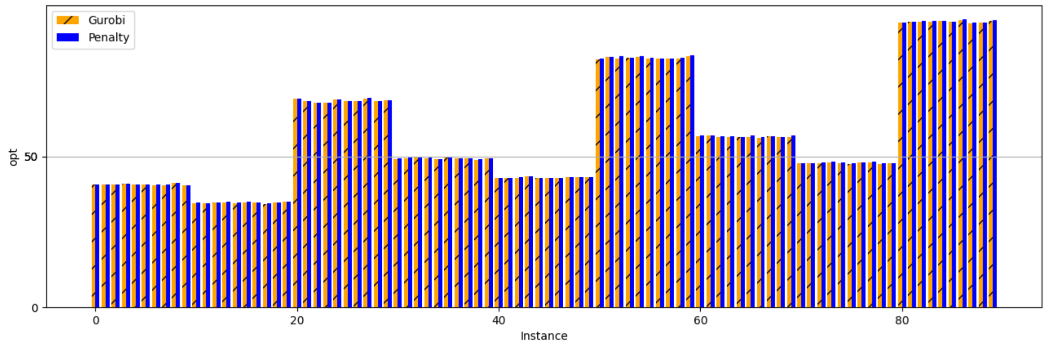
For the numerical test, we use an instance of 90 generated covariance matrices Q and return vectors μ from dimensions $n = 200$, $n = 300$ and $n = 400$ provided by Frangioni and Gentile[1]. We used Gurobi[2] to determine a best possible value of each test problem. Gurobi was set to run for at least 60 seconds in all of the 90 cases and we compared the results to our exact penalty method. The computation were carried out via Python, where we chose $\alpha_0 = \rho = \beta = 1$. The spectral gradient was run for 1000 iterations or until a stationarity to the tolerance of 10^{-4} was reached. In the outer loop of the exact penalty method, we set $\alpha_{k+1} = 2\alpha_k$ until a complementarity $\|s^k \circ y^k\|_\infty < 10^{-3}$ was reached. On average, the required computation time for a full run of the exact penalty method for a single instance was

$$t_{200} = 3.94s, \quad t_{300} = 5.69s, \quad t_{400} = 6.66s,$$

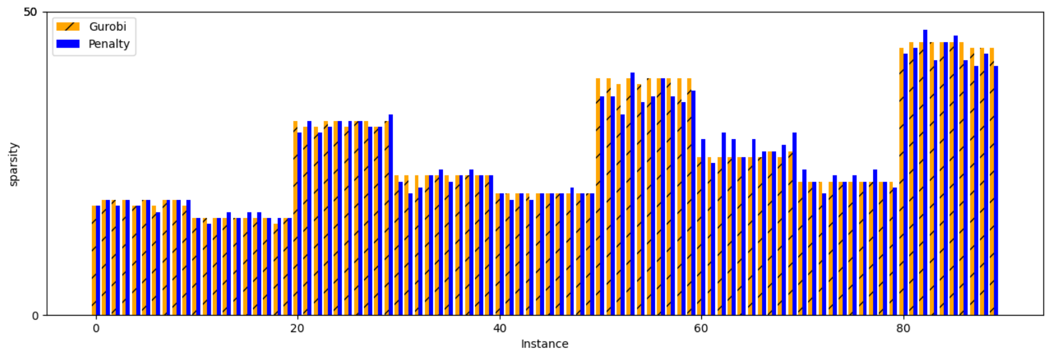
for dimensions 200 to 400, respectively (Figure 1).

[1] http://groups.di.unipi.it/optimize/Data/MV.html

[2] https://www.gurobi.com/



(a) Optimal values of $f(x) + \|x\|_0$ reached by Gurobi and the penalty approach



(b) Sparsity reached by Gurobi and the penalty approach

Figure 1: Overview of the sparse portfolio tests.

5.2 Sparse Dictionary Learning

5.2.1 Problem Formulation

We consider the matrix-valued optimization problem

$$\min_{C,D} \frac{1}{2} \left\| D^T C - Z \right\|_F^2 + \rho \left\| C \right\|_0 \quad \text{s.t.} \quad \|d_i^T\|_2 \leq 1,$$

where $D \in \mathbb{R}^{l \times n}$ is considered a dictionary with rows d_i^T from which a couple of entries are drawn to create matrix $Z \in \mathbb{R}^{n \times m}$. This process is represented by multiplying D^T with a suitable sparse matrix C, which gives rise to the formulation with the sparsity inducing ℓ_0-norm regularization term. Note that, in this context, we use

$$\left\| C \right\|_0 = \left\| \text{vec}(C) \right\|_0 .$$

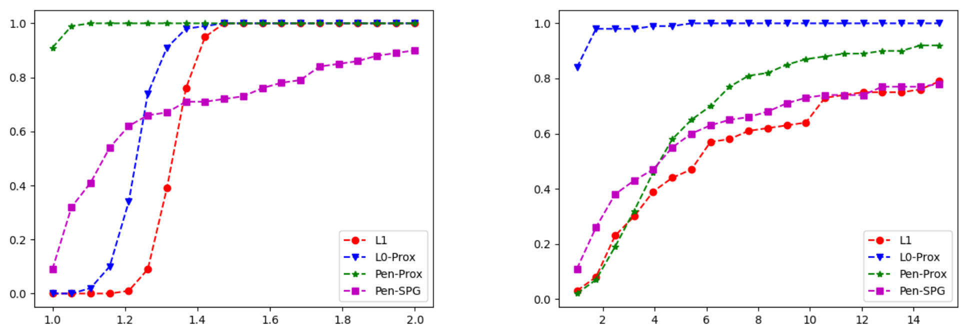
For this class of problems, the feasible set does not give any restrictions for C, the variable we impose sparsity on, therefore projection onto the epigraph of the componentwise absolute value function as well as the proximal operator prox_γ^{SP} can be computed in a straightforward manner. Since replacing the ℓ_0-norm by the ℓ_1-norm is a popular approach that turns the objective function into a biconvex optimization problem, we compare in the following the results achieved by a hard-thresholding approach, using the proximal operator of the ℓ_0-norm to a soft-thresholding approach, using the proximal operator of the ℓ_1-norm, and finally the exact penalty with spectral gradient and the proximal point method.

5.2.2 Numerical Tests

We copy the setup from [9] and create 100 test problems, with $n = 100$, $l = 200$ and $m = 300$. The entries of C^0 and D^0 are drawn from a standard normal distribution $\mathcal{N}(0, 1)$, projected onto the feasible set. Again, for simplicity, we set $\rho = 1$. In case of the exact penalty method, the inner solver was set to run for 10^4 iterations or a until a stationarity of tolerance 10^{-5} was reached. We also set $\alpha_0 = 1$ and $\alpha_{k+1} = 1.5\,\alpha_k$ until

$$\langle |C^k|, Y^k \rangle_F := \text{trace}((C^k)^T Y^k) \leq 10^{-3}.$$

In case of the hard and soft thresholding method, we set the maximum number of iterations to 10^5 and the tolerance of stationarity to 10^{-6}. We draw the performance profiles regarding the optimal function values reached and the required computational time. As it turns out, the exact



(a) Performance profile on the optimal target value.

(b) Performance profile on the computation time.

Figure 2: Overview of the sparse portfolio tests.

penalty method with the inner proximal point solver (Pen-Prox) has a chance of approximately 90% of achieving the best possible value, while the spectral gradient (Pen-SPG) was able to find the best possible value in 10% of the cases (Figure 2a). Both thresholding methods were never able to find the best possible solution. The spectral gradient manages to stay in the range of the optimal solution for approximately 60% of all cases and is then outperformed by the solver ℓ_0 and ℓ_1.

Considering computational time, the ℓ_0 solver wins the race, which was to be expected. Surprisingly enough, the penalty approaches stay well within the performance of the soft-thresholding operator, with the Pen-Prox overtaking and landing at 80% chance to require roughly ten times the computational time of the hard-thresholding solver (Figure 2b).

5.3 Sparse Untargeted Adversarial Attacks on Neural Networks

5.3.1 Problem Formulation

Let $M := \{(x^d, y^d) \in \mathcal{X} \times \mathcal{Y},\ d = 1, ..., S\}$ denote a set of data samples, with sample space $\mathcal{X} \subset \mathbb{R}^n$ and label space $\mathcal{Y} \subset \mathbb{R}^m$. A neural network is a function $\phi : \mathbb{R}^n \to \mathbb{R}^m$, which aims to satisfy

$$P\left(\phi(x^d)\right) = y^d, \quad \forall d = 1, ..., S,$$

where ϕ is a composition of several functions

$$\phi = f_l \circ f_{l-1} \circ \cdots \circ f_1,$$

each corresponding to a layer $k = 1, ..., l$, comprised of a weight matrix W^k, a bias b^k and some activation function $\sigma_k : \mathbb{R}^{n_k} \to \mathbb{R}^{m_k}$ so that essentially $f_k(x) = \sigma_k(W^k f_{k-1}(x) + b^k)$ and $f_0 = \text{id}$. The function $P(\cdot)$ processes the vector of logits (the raw output of network ϕ) in a suitable way. Training of a neural network is carried out by the minimization of some loss function L across the data samples in M, in the sense that one has to solve

$$\min_{\{(W^k,b^k),\ k=1,\dots,l\}} \frac{1}{S} \sum_{d=1}^{S} L\left(P(\phi(x^d)), y^d\right).$$

Assume that y^d is a probability mass function, which belongs to some discrete distribution, that is $\mathcal{Y} = \{y \in \mathbb{R}^m \ :\ y \geq 0,\ \|y\|_1 = 1\}$. One chooses P as a function to carry the logits into the probability space $P(\phi(x^d)) = p^d$, for instance, by the softmax

$$p^d = \text{softmax}(\phi(x^d)) := \frac{z^d}{\|z^d\|_1}, \quad \text{with} \quad z_i^d = \exp\left(\phi(x^d)_i\right),$$

to then minimize a statistical distance between y^d and p^d, where now p_i^d measures the probability of x^d belonging to class $i \in \{1, ..., m\}$. For this purpose, one usually considers the Kullback-Leibler divergence

$$KL(y^d, p^d) := \sum_{i=1}^{m} y_i^d \log\left(\frac{y_i^d}{p_i^d}\right),$$

and, since y^d is fixed, we may simply choose $L(P(\phi(x^d)), y^d) := -\sum_{i=1}^{m} y_i^d \log(p_i^d)$, which is commonly called the *cross entropy loss*. We now assume y^d to be one-hot encoded, which means that $y^d = e_{i_d}$, i.e. y^d is the i_d-th unit vector of $\mathbb{R}^m$. The model ϕ should then satisfy

$$\underset{i=1,\dots m}{\text{argmax}}\ P(\phi(x^d)) = i_d.$$

In the sparse adversarial attack setting, the model ϕ has already been trained and we try to find a suitable deviation δ, where $\|\delta\|_0$ is small, to throw off the previous classification of an input x^d, that is

$$\underset{i=1,\dots m}{\text{argmax}}\ P(\phi(x^d + \delta)) \neq \underset{i=1,\dots,m}{\text{argmax}}\ P(\phi(x^d)).$$

Since the latter cannot be chosen as a constraint, we instead want to minimize the auxiliary function

$$F(\delta) = \log(p_{i_d}^d(\delta)) - \sum_{\substack{j=1 \\ j \neq i_d}}^{m} \log(p_j^d(\delta)), \quad \text{with} \quad p^d(\delta) = P(\phi(x^d + \delta)).$$

This reflects the idea that, by minimizing the cross entropy loss, $\log(p_{i_d}^d)$ has to have been large, since i_d was the predicted label, whereas the logarithm with respect to the other classes $\{1, ..., n\} \setminus \{i_d\}$, must have been small. This, of course, is simply a heuristic approach, as it is, for instance, unclear whether an appropriate scaling between the two summands in F should be in place. We now try to find solutions δ^ρ of

$$\min_{\delta} \frac{1}{\rho} F(\delta) + \|\delta\|_0, \quad \text{s.t.} \quad x + \delta \in \mathcal{X}$$

for decreasing values of ρ until for a solution δ^ρ, the input $x^d + \delta^\rho$ is classified differently from x^d. This approach was presented in [6], albeit with a different selection of penalty functions F and, additionally, the authors of the aforementioned work focused on targeted network attacks, where a certain class was given, x^d should be recognized as. This might in some cases lead to better results, since a model tasked with discerning, for instance, numbers might confuse a one with a seven easier than a one with an eight.

5.3.2 Numerical Test

We execute the adversarial attacks on the well known MNIST data set, which consist of 60000 hand-drawn black and white images with 28×28 pixels, that is $\mathcal{X} = [0,1]^{28\times 28}$, depicting numbers from 0 to 9. The set of labels can therefore be represented as $\mathcal{Y} = \{0,1\}^{10}$.

We set up a neural network of the forward structure as specified in Table 1. We chose a dropout of 0.2 and activated the convolution and fully connected layers with the sigmoid function each.

Convolution 2D layer with 32 filters and kernel size 3
Convolution 2D layer with 32 filters and kernel size 3
Average Pooling 2D layer of size 2×2
Convolution 2D layer with 64 filters and kernel size 3
Convolution 2D layer with 64 filters and kernel size 3
Average Pooling 2D layer of size 2×2
Fully connected layer of size 200
Fully connected layer of size 200
Dense output layer of size 10

Table 1: Layers of the MNIST Model

The training was carried out via Tensorflow[3], which is freely available as a python package, and computation was done on a Nvidia RTX 3070 GPU, where we attained an accuracy of 99.92% and a loss of 0.0022 across the training set, and an accuracy of 99.22% and a loss of 0.0352 on the test set.

We deployed an adversarial attack on the first 100 data samples of the training set. To this end, we chose $\rho^0 = 10$ and $\rho_{k+1} = 0.9 \cdot \rho_k$ until a solution δ^{ρ_k} of

$$\min_{\delta} \ \frac{1}{\rho_k} F(\delta) + \|\delta\|_0 , \quad \text{s.t.} \quad 0 \le x^d + \delta \le 1$$

would satisfy

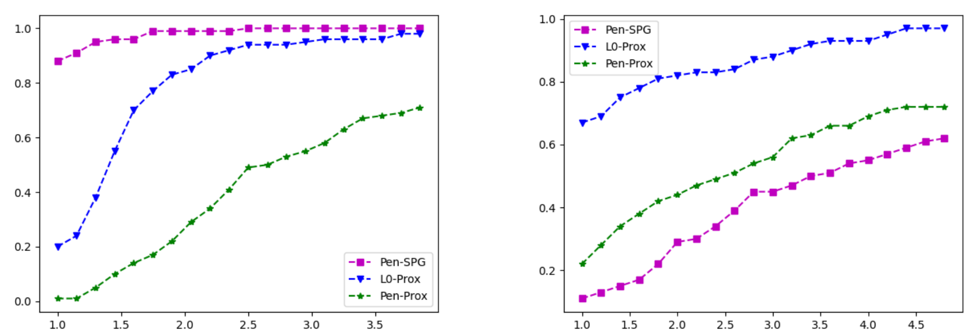
$$\operatorname*{argmax}_{i=1,\dots m} P(\phi(x^d + \delta)) \neq \operatorname*{argmax}_{i=1,\dots,m} P(\phi(x^d)).$$

The problems were solved with, on the one hand, a straightforward application of the ℓ_0 proximal gradient method (L0-Prox) and, on the other hand, the exact penalty method with the spectral gradient inner solver (Pen-SPG) and proximal inner solver (Pen-Prox). To compute ∇F, we

[3] https://www.tensorflow.org/

rely on the automatic differentiation functionality of Tensorflow to obtain the gradient of ϕ with respect to the input $x^d + \delta$. The exact penalty parameter was initialized to $\alpha^0 = 1.0$ and increased in increments of 10 until complementarity was reached to a tolerance of 10^{-3}. The inner solvers were set to run for 10^3 iterations until either stationarity of tolerance 10^{-4} was reached or the algorithm did no longer measure any progress in the attained function values over 10 iterations.

Notice that we had to slightly adapt projections and prox-operators to incorporate the projection of δ onto $\delta_{ij} \in [-x_{ij}^d, 1 - x_{ij}^d]$. It turns out that Pen-SPG behave significantly better than Pen-Prox, while also beating the hard-threshold approach L0-Prox in most of the cases. Pen-SPG generated adversarial images with on average $\|\delta\|_0 \approx 11$, whereas L0-Prox generated adversarial images with $\|\delta\|_0 \approx 17$ and Pen-Prox coming in last with $\|\delta\|_0 \approx 29$. The performance profile on the attained sparsity of δ and the required computational time is detailed in Figure 3a and Figure 3b. Not surprisingly, the hard thresholding operator is, again, significantly ahead in terms of computational time.



(a) Performance profile on the optimal target value.

(b) Performance profile on the computation time.

Figure 3: Overview of the adversarial attack tests.

6 Final Remarks

This paper generalizes the authors' previous work [17] and presents an exact penalty technique for sparse optimization problems where the variables are not necessarily sign-constrained. The theoretical results provide a very strong relation between the exact penalty approach and the given sparse optimization problem, even in terms of stationary points, and the numerical results look very promising.

In contrast to [17], however, the newly introduced penalty function is nonsmooth like most exact penalty functions for standard nonlinear programs. On the other had, it is known that there exist differentiable exact penalty techniques, see, e.g., the survey article [10]. These differentiable exact penalty methods have a very strong theoretical background, but their practical application is limited due to the fact that their evaluation is, in general, very expensive. In our particular reformulation of the sparsity term, however, the resulting constraints have a very simple structure, in which case the differentiable exact penalty functions might be implemented in a much more efficient way than for general constraints. We therefore leave it as part of our future research to investigate the application of differentiable exact penalty functions to our particular reformulation of sparse optimization problems.

Statements

Competing Interests

There are no competing interests to report.

Data Availability

We summarize the data availability for each of the considered test problems in the list below.

1. The mixed-integer portfolio optimization problems by Frangioni and Gentile are publicly available (c.f. footnote 1). To our knowledge, the problems are randomly generated. The specific instance for this manuscript lies with the authors and is available on request.

2. The sparse dictionary learning instances have been generated procedurally and randomly from a fixed random seed with the NumPy[4] package available for Python.

3. The MNIST data set for the considered adversarial attack tests is publicly available and, in our case, was downloaded directly from TensorFlow. The CNN used for the classification task was trained as stated and saved as keras file. It is available on request.

References

[1] R. Andreani, G. Haeser, R. Prado, and L. Secchin. Primal-dual global convergence of an augmented Lagrangian method under the error bound condition. *Optimization Online*, 2025. https://optimization-online.org/?p=31199.

[2] M. Benko, M. Červinka, and T. Hoheisel. Sufficient conditions for metric subregularity of constraint systems with applications to disjunctive and ortho-disjunctive programs. *Set-Valued and Variational Analysis*, 30(1):143–177, 2021. doi:10.1007/s11228-020-00569-7.

[3] D. Bienstock. Computational study of a family of mixed-integer quadratic programming problems. *Mathematical Programming*, 74:121–140, 1996. doi:10.1007/BF02592208.

[4] E. G. Birgin and J. M. Martínez. *Practical Augmented Lagrangian Methods for Constrained Optimization*. SIAM, May 2014. doi:10.1137/1.9781611973365.

[5] O. P. Burdakov, C. Kanzow, and A. Schwartz. Mathematical programs with cardinality constraints: reformulation by complementarity-type conditions and a regularization method. *SIAM Journal on Optimization*, 26(1):397–425, 2016. doi:10.1137/140978077.

[6] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, page 39–57. IEEE, 2017. doi:10.1109/sp.2017.49.

[4]https://numpy.org/

[7] F. H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990. doi:10.1137/1.9781611971309.

[8] Y. Cui and J.-S. Pang. *Modern Nonconvex Nondifferentiable Optimization*. SIAM, 2021. doi:10.1137/1.9781611976748.ch4.

[9] A. De Marchi. Proximal gradient methods beyond monotony. *Journal of Nonsmooth Analysis and Optimization*, Volume 4 (Original research articles), 2023. doi:10.46298/jnsao-2023-10290.

[10] G. Di Pillo. Exact penalty methods. E. Spedicato, ed. algorithms for continuous optimization. NATO ASI Series Vol. 434, 1994. doi:10.1007/978-94-009-0369-2_8.

[11] M. Durea and R. Strugariu. *An Introduction to Nonlinear Optimization Theory*. De Gruyter, 2014. doi:10.2478/9783110426045.

[12] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. doi:10.1198/016214501753382273.

[13] M. Feng, J. E. Mitchell, J.-S. Pang, X. Shen, and A. Wächter. Complementarity formulations of ℓ_0-norm optimization problems. *Pacific Journal of Optimization*, 14(2):273–305, 2018.

[14] H. Gfrerer. On directional metric subregularity and second-order optimality conditions for a class of nonsmooth mathematical programs. *SIAM Journal on Optimization*, 23(1):632–665, 2013. doi:10.1137/120891216.

[15] J.-y. Gotoh, A. Takeda, and K. Tono. DC formulations and algorithms for sparse optimization problems. *Mathematical Programming*, 169(1):141–176, 2017. doi:10.1007/s10107-017-1181-0.

[16] C. Kanzow, A. Schwartz, and F. Weiß. The sparse(st) optimization problem: reformulations, optimality, stationarity, and numerical results. *Computational Optimization and Applications*, 90(1):77–112, 2024. doi:10.1007/s10589-024-00625-0.

[17] C. Kanzow and F. Weiß. An exact penalty approach for general ℓ_0-sparse optimization problems, 2023. doi:10.48550/ARXIV.2312.15706.

[18] H. Y. Le. Generalized subdifferentials of the rank function. *Optimization Letters*, 7(4):731–743, 2012. doi:10.1007/s11590-012-0456-x.

[19] H. A. Le Thi, T. P. Dinh, H. M. Le, and X. T. Vo. DC approximation approaches for sparse optimization. *European Journal of Operational Research*, 244(1):26–46, 2015. doi:10.1016/j.ejor.2014.11.031.

[20] Y.-C. Liang and J. J. Ye. Optimality conditions and exact penalty for mathematical programs with switching constraints. *Journal of Optimization Theory and Applications*, 190(1):1–31, 2021. doi:10.1007/s10957-021-01879-y.

[21] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77, 1952. doi:10.2307/2975974.

[22] L. Minchenko and A. Tarakanov. On error bounds for quasinormal programs. *Journal of Optimization Theory and Applications*, 148(3):571–579, 2011. doi:10.1007/s10957-010-9768-0.

[23] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation I.* Springer Berlin Heidelberg, 2006. doi:10.1007/3-540-31247-1.

[24] B. S. Mordukhovich. *Variational Analysis and Applications.* Springer Cham, 2018. doi:10.1007/978-3-319-92775-6.

[25] T. B. T. Nguyen, H. A. L. Thi, H. M. Le, and X. T. Vo. DC approximation approach for ℓ_0-minimization in compressed sensing. In *Advanced Computational Methods for Knowledge Engineering: Proceedings of 3rd International Conference on Computer Science, Applied Mathematics and Applications-ICCSAMA 2015*, pages 37–48. Springer, 2015. doi:10.1007/978-3-319-17996-4_4.

[26] 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 *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–4, 2019. doi:10.1109/SampTA45681.2019.9030937.

[27] D. Ralph and S. J. Wright. Some properties of regularization and penalization schemes for MPECs. *Optimization Methods and Software*, 19(5):527–556, 2004. doi:10.1080/10556780410001709439.

[28] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis.* Springer Berlin Heidelberg, 1998. doi:10.1007/978-3-642-02431-3.

[29] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 2018. doi:10.1111/j.2517-6161.1996.tb02080.x.

[30] A. M. Tillmann, D. Bienstock, A. Lodi, and A. Schwartz. Cardinality minimization, constraints, and regularization: a survey. *SIAM Review*, 66(3):403–477, 2024. doi:10.1137/21M142770X.

[31] J. J. Ye. The exact penalty principle. *Nonlinear Analysis: Theory, Methods & Applications*, 75(3):1642–1654, 2012. doi:https://doi.org/10.1016/j.na.2011.03.025.

[32] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2), 2010. doi:10.1214/09-aos729.

[33] S. Zhang and J. Xin. Minimization of transformed ℓ_1 penalty: theory, difference of convex function algorithm, and robust application in compressed sensing. *Mathematical Programming*, 169(1):307–336, 2018. doi:10.1007/s10107-018-1236-x.