Journal of Optimization Theory and Applications (2024) 202:795–833
https://doi.org/10.1007/s10957-024-02448-9

Exterior-Point Optimization for Sparse and Low-Rank Optimization

Shuvomoy Das Gupta[1] · **Bartolomeo Stellato**[2] · **Bart P. G. Van Parys**[3]

Received: 28 June 2023 / Accepted: 26 April 2024 / Published online: 26 May 2024

Abstract

Many problems of substantial current interest in machine learning, statistics, and data science can be formulated as sparse and low-rank optimization problems. In this paper, we present the nonconvex exterior-point optimization solver (NExOS)—a first-order algorithm tailored to sparse and low-rank optimization problems. We consider the problem of minimizing a convex function over a nonconvex constraint set, where the set can be decomposed as the intersection of a compact convex set and a nonconvex set involving sparse or low-rank constraints. Unlike the convex relaxation approaches, NExOS finds a locally optimal point of the original problem by solving a sequence of penalized problems with strictly decreasing penalty parameters by exploiting the nonconvex geometry. NExOS solves each penalized problem by applying a first-order algorithm, which converges linearly to a local minimum of the corresponding penalized formulation under regularity conditions. Furthermore, the local minima of the penalized problems converge to a local minimum of the original problem as the penalty parameter goes to zero. We then implement and test NExOS on many instances from a wide variety of sparse and low-rank optimization problems, empirically demonstrating that our algorithm outperforms specialized methods.

Keywords Nonconvex optimization · Sparse optimization · Low-rank optimization · First-order algorithms

Communicated by Clément W. Royer.

✉ Bartolomeo Stellato
bstellato@princeton.edu

Shuvomoy Das Gupta
sdgupta@mit.edu

Bart P. G. Van Parys
vanparys@mit.edu

[1] Operations Research Center, Massachusetts Institute of Technology, Cambridge, USA

[2] Department of Operations Research and Financial Engineering, Princeton University, Princeton , USA

[3] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, USA

Mathematics Subject Classification 65K05 · 90C30

1 Introduction

This paper studies optimization problems involving a strongly convex and smooth cost function over a closed nonconvex constraint set $\mathcal{X}$ involving sparse or low-rank constraints. We propose a first-order algorithm *nonconvex exterior-point optimization solver* (NExOS) to solve such problems numerically. We can write such problems as:

$$\begin{array}{ll}\text{minimize} & f(x) + (\beta/2)\|x\|^2 \\ \text{subject to} & x \in \mathcal{X},\end{array} \tag{$\mathcal{P}$}$$

where x takes value in a finite-dimensional vector space $\mathbf{E}$ over the reals, f is a strongly convex and smooth function. In Appendix B.1, we generalize our framework to the case when f is non-smooth convex.

The regularization parameter $\beta > 0$ is commonly introduced in statistics and machine learning problems to reduce the generalization error without increasing the training error [33, §5.2.2]. In this paper, there is also a theoretical consideration behind including the term $\frac{\beta}{2}\|x\|^2$ in problem ($\mathcal{P}$). NExOS finds a locally optimal point of problem ($\mathcal{P}$) by solving a sequence of penalized subproblems with strictly decreasing penalty parameters, where each penalized subproblem is solved by a first-order algorithm. Under the presence of $\frac{\beta}{2}\|x\|^2$ with $\beta > 0$, we can prove that each penalized subproblem is locally strongly convex and smooth admitting a unique local minimum (see Proposition 1), which in turn ensure linear convergence of the first-order method to that local minimum (see Theorem 1). In the Numerical Experiments section, we demonstrate that β can be set to a value as small as 10^{-8}. This empirical evidence suggests that, in practice, the impact of β on the objective value can be made negligible, yet one can still reap the theoretical benefits. Therefore, while β plays a crucial role in the theoretical aspects of our algorithm, its influence on the problems considered in the Numerical Experiments section is minimal and can be adjusted as per the problem's requirements.

Furthermore, $\mathbf{E}$ is equipped with inner product $\langle \cdot \mid \cdot \rangle$ and norm $\|\cdot\| = \sqrt{\langle x \mid x \rangle}$. For $\mathbf{E} = \mathbf{R}^d$, we have $\langle x \mid y \rangle = x^\top y$ for $x, y \in \mathbf{R}^d$, and for $\mathbf{E} = \mathbf{R}^{m\times n}$, we have $\langle X \mid Y \rangle = \mathbf{tr}(X^\top Y)$, for $X, Y \in \mathbf{R}^{m\times n}$. The constraint set $\mathcal{X}$ is closed and nonconvex and can be decomposed as the intersection of a compact convex set and a nonconvex set involving sparse or low-rank constraints. Sparse and low-rank constraint sets are very important in modeling many machine learning problems, because they allow for high interpretability, speed-ups in computation, and reduced memory requirements [40].

Sparsity-constrained optimization Sparsity constraints have found applications in many practical settings, e.g., gene expression analysis [38, pp. 2–4], sparse regression [40, pp. 155–157], signal transmission and recovery [22, 67], hierarchical sparse polynomial regression [14], and best subset selection [13], just to name a few. In these problems, the constraint set $\mathcal{X}$ decomposes as $\mathcal{X} = \mathcal{C} \bigcap \mathcal{N}$, where $\mathcal{C}$ is a compact convex set, and

$$\mathcal{N} = \{x \in \mathbf{R}^d \mid \mathbf{card}(x) \leq k\}, \tag{1}$$

where $\mathbf{card}(x)$ counts the number of nonzero elements in x. In these optimization problems, $\mathcal{C}$ can be a polyhedron, infinity-norm ball, box constraint set, or probability simplex; these sets usually show up in applications involving econometrics, housing price prediction, air-quality prediction, signal processing, and meteorology [8, 11, 15, 25].

In this paper, we apply NExOS to solve the sparse regression problem for both synthetic and real-world datasets in Sect. 4.1, which is concerned with approximating a vector $b \in \mathbf{R}^m$ with a linear combination of at most k columns of a matrix $A \in \mathbf{R}^{m\times d}$ with bounded coefficients. This problem has the form:

$$\begin{array}{l}\text{minimize } \|Ax - b\|_2^2 + (\beta/2)\|x\|_2^2 \\ \text{subject to } \mathbf{card}(x) \leq k, \quad \|x\|_\infty \leq \Gamma,\end{array} \tag{SR}$$

where $x \in \mathbf{R}^d$ is the decision variable, and $A \in \mathbf{R}^{m\times d}$, $b \in \mathbf{R}^m$, and $\Gamma > 0$ are problem data.

Low-rank optimization We can write low-rank optimization problems in the form of problem ($\mathcal{P}$), which are common in machine learning applications such as collaborative filtering [40, pp. 279–281], design of online recommendation systems [21, 45], bandit optimization [41], data compression [34, 46, 60], and low rank kernel learning [2]. In these applications, the constraint set $\mathcal{X}$ decomposes as $\mathcal{X} = \mathcal{C} \bigcap \mathcal{N}$, where $\mathcal{C}$ is a compact convex set, and

$$\mathcal{N} = \{X \in \mathbf{R}^{m\times d} \mid \mathbf{rank}(X) \leq r\}. \tag{2}$$

In these optimization problems, $\mathcal{C}$ can be matrix-norm ball, Frobenius-norm ball, hyperplane/half-space induced by trace [9, 10]. In this paper, we apply NExOS to solve the affine rank minimization problem:

$$\begin{array}{l}\text{minimize } \|\mathcal{A}(X) - b\|_2^2 + (\beta/2)\|X\|_F^2 \\ \text{subject to } \mathbf{rank}(X) \leq r, \quad \|X\|_2 \leq \Gamma,\end{array} \tag{RM}$$

where $X \in \mathbf{R}^{m\times d}$ is the decision variable, $b \in \mathbf{R}^k$ is noisy measurement data, and $\mathcal{A} : \mathbf{R}^{m\times d} \to \mathbf{R}^k$ is a linear map. The parameter $\Gamma > 0$ is the upper bound for the spectral norm of X. The affine map $\mathcal{A}$ is determined by k matrices $A_1, \ldots, A_k$ in $\mathbf{R}^{m\times d}$ where $\mathcal{A}(X) = (\mathbf{tr}(A_1^T X), \ldots, \mathbf{tr}(A_k^T X))$. We present several numerical experiments to solve (RM) using NExOS for both synthetic and real-world datasets in Sect. 4.2.

1.1 Related Work

Convex relaxation approach Due to the presence of the nonconvex set $\mathcal{X}$, the nonconvex problem ($\mathcal{P}$) is $\mathcal{NP}$-hard [35]. A common way to deal with this issue is to avoid this inherent nonconvexity altogether by convexifying the original problem. The relaxation of the sparsity constraint leads to the popular LASSO formulation and its variants

[38], whereas relaxation of the low-rank constraints produces the nuclear norm based convex models [28].

The basic advantage of the convex relaxation technique is that, in general, a globally optimal solution to a convex problem can be computed reliably and efficiently [20, §1.1], whereas for nonconvex problems a local optimal solution is often the best one can hope for. Furthermore, if certain statistical assumptions on the data generating process hold, then it is possible to recover exact solutions to the original nonconvex problems with high probability by solving the convex relaxations (see [38] and the references therein).

However, when stringent assumptions do not hold, then solutions to the convex formulations can be of poor quality and may not scale very well [40, §6.3 and §7.8]. In this situation, the nonconvexity of the original problem must be confronted directly, because such nonconvex formulations capture the underlying problem structures more accurately than their convex counterparts.

First-order methods To that goal, first-order algorithms such as hard thresholding algorithms, e.g., IHT [17], NIHT [18], HTP [30], CGIHT [16], address nonconvexity in sparse and low-rank optimization by implementing variants of projected gradient descent with projection taken onto the sparse and/or low-rank set.

While these first-order methods have been successful in recovering low-rank and sparse solutions in underdetermined linear systems, they too require assumptions on the data such as the *restricted isometry property* for recovering true solutions [40, §7.5]. Furthermore, to converge to a local minimum, hard thresholding algorithms require the spectral norm of the measurement matrix to be less than one, which is a restrictive condition [17].

Besides hard thresholding algorithms, heuristics based on first-order algorithms such as the alternating direction method of multipliers (ADMM) have gained a lot of traction in the last few years. Though ADMM was originally designed to solve convex optimization problems, since the idea of implementing this algorithm as a general purpose heuristic to solve nonconvex optimization problems was introduced in [19, §9.1−9.2], ADMM-based heuristics have been applied successfully to approximately solve nonconvex problems in many different application areas [25, 63].

However, the biggest drawback of these heuristics based on first-order methods comes from the fact that they take an algorithm designed to solve convex problems and apply it verbatim to a nonconvex setup. As a result, these algorithms often fail to converge, and even when they do, it need not be a local minimum, let alone a global one [62, §2.2]. Also, empirical evidence suggests that the iterates of these algorithms may diverge even if they come arbitrarily close to a locally optimal solution during some iteration. The main reason is that these heuristics do not establish a clear relationship between the local minimum of problem ($\mathcal{P}$) and the fixed point set of the underlying operator that controls the iteration scheme. An alternative approach that has been quite successful empirically in finding low-rank solutions is to consider an unconstrained problem with Frobenius norm penalty and then using alternating minimization to compute a solution [68]. However, the alternating minimization approach may not converge to a solution and should be considered a heuristic [68, §2.4].

Discrete optimization approach For these reasons above, in the last few years, there has been significant interest in addressing the nonconvexity present in many optimization problems directly via a discrete optimization approach. In this way, a particular nonconvex optimization problem is formulated exactly using discrete optimization techniques and then specialized algorithms are developed to find a certifiably optimal solution. This approach has found considerable success in solving machine learning problems with sparse and low-rank optimization [12, 66]. A mixed integer optimization approach to compute near-optimal solutions for sparse regression problem, where problem dimension $d = 1000$, is computed in [13]. In [15], the authors propose a cutting plane method for a similar problem, which works well with mild sample correlations and a sufficiently large dimension. In [39], the authors design and implement fast algorithms based on coordinate descent and local combinatorial optimization to solve sparse regression problem with a three-fold speedup where $d \approx 10^6$. In [10], the authors propose a framework for modeling and solving low-rank optimization problems to certifiable optimality via symmetric projection matrices.

However, the runtime of these discrete optimization based algorithms can often become prohibitively long as the problem dimensions grow [13]. Also, these discrete optimization algorithms have efficient implementations only for a narrow class of loss functions and constraint sets; they do not generalize well if a minor modification is made to the problem structure, and in such a case they often fail to find a solution point in a reasonable amount of time even for smaller dimensions [12]. Furthermore, one often relies on commercial softwares, such as `Gurobi`, `Mosek`, or `Cplex` to solve these discrete optimization problems, thus making the solution process somewhat opaque [13, 66].

1.2 Contributions

The main contribution of this work is to propose NExOS: a first-order algorithm tailored for nonconvex optimization problems of the form ($\mathcal{P}$). The term *exterior-point* originates from the fact that the iterates approach a local minimum from outside of the feasible region; it is inspired by the convex exterior-point method first proposed by Fiacco and McCormick in the 1960s [29, §4]. By exploiting the underlying geometry of the constraint set, we construct an iterative method that finds a locally optimal point of the original problem via an outer loop consisting of increasingly accurate penalized formulations of the original problem by reducing only one penalty parameter. Each penalized problem is then solved by applying an inner algorithm that implements a variant of the Douglas–Rachford splitting algorithm.

We prove that NExOS, besides avoiding the drawbacks of convex relaxation and discrete optimization approach, has the following favorable features. First, the penalized problem has strong convexity and smoothness around local minima, but can be made arbitrarily close to the original nonconvex problem by reducing the penalty parameter. Second, under mild regularity conditions, the inner algorithm finds local minima for the penalized problems at a linear convergence rate, and as the penalty parameter goes to zero, the local minima of the penalized problems converge to a local minimum of the original problem. Furthermore, we show that, when those regularity conditions

do not hold, the inner algorithm is still guaranteed to subsequentially converge to a first-order stationary point of the penalized problem at the rate $o(1/\sqrt{k})$.

We implement NExOS in the open-source `Julia` package `NExOS.jl` and test it extensively on many synthetic and real-world instances of different nonconvex optimization problems of substantial current interest. We demonstrate that NExOS very quickly computes solutions that are competitive with or better than specialized algorithms on various performance measures. `NExOS.jl` is available at https://github.com/Shuvomoy/NExOS.jl.

Organization of the paper

The rest of the paper is organized as follows. We describe our NExOS framework in Sect. 2. We provide convergence analysis of the algorithm in Sect. 3. Then we demonstrate the performance of our algorithm on several nonconvex optimization problems of significant current interest in Sect. 4. The concluding remarks are presented in Sect. 5.

2 Our Approach

The backbone of our approach is to address the nonconvexity by working with an asymptotically exact nonconvex penalization of problem ($\mathcal{P}$), which enjoys local convexity around local minima. We use the notation $\iota_{\mathcal{X}}(x)$ that denotes the indicator function of the set $\mathcal{X}$ at x, which is 0 if $x \in \mathcal{X}$ and ∞ else. Using this, we can write problem ($\mathcal{P}$) as an unconstrained optimization problem, where the objective is $f(x) + (\beta/2)\|x\|^2 + \iota_{\mathcal{X}}(x)$. In our penalization, we replace the indicator function $\iota_{\mathcal{X}}$ with its *Moreau envelope* with positive parameter μ:

$$ {}^{\mu}\iota(x) = \min_{y} \{\iota_{\mathcal{X}}(y) + (1/2\mu)\|y - x\|^2\} = (1/2\mu)d^2(x), \tag{3} $$

where $d(x)$ is the Euclidean distance of the point x from the set $\mathcal{X}$.

Properties of Moreau envelope for a nonconvexset.

The function ${}^{\mu}\iota$, though nonconvex, has many desirable attributes that greatly simplify design and convergence analysis of our algorithm. We summarize these properties below; See [3, Proposition 12.9] for the first four properties, and Proposition 1 in Sect. 3 for the last one.

1. *Bounded.* The function ${}^{\mu}\iota$ is bounded on every compact set. In contrast, $\iota_{\mathcal{X}}$ is an extended valued function that takes the value $+\infty$ outside the set $\mathcal{X}$.
2. *Finite and jointly continuous.* For every $\mu > 0$ and $x \in \mathbf{E}$, the function ${}^{\mu}\iota(x)$ is jointly continuous and finite. Therefore, ${}^{\mu}\iota$ is continuous on $\mathbf{E}$. In contrast, the indicator function $\iota_{\mathcal{X}}$ is not continuous.
3. *Accuracy of approximation controlled by μ.* With decreasing μ, the approximation ${}^{\mu}\iota$ monotonically increases to $\iota_{\mathcal{X}}$, i.e., for any positive μ_1, μ_2 such that $0 \leq \mu_1 \leq \mu_2$, we have

$$ 0 \leq {}^{\mu_2}\iota_{\mathcal{X}}(x) \leq {}^{\mu_1}\iota_{\mathcal{X}}(x) \leq \iota_{\mathcal{X}}(x) $$

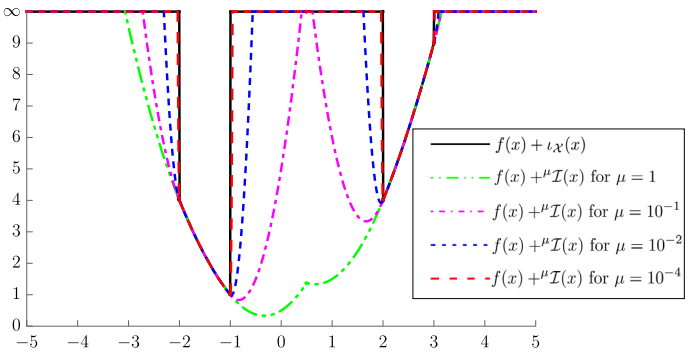


Fig. 1 An illustration of how the penalized cost function in problem ($\mathcal{P}_\mu$) compares against the original cost function in problem ($\mathcal{P}$) for different values of μ. Note that the regularization parameter β is kept fixed at its initial value 1 throughout

for any $x \in \mathbf{E}$.

4. *Asymptotically equal to* $\iota_{\mathcal{X}}$. The approximation ${}^{\mu}\iota$ is asymptotically equal to $\iota_{\mathcal{X}}$ as μ goes to zero, i.e., we have the point-wise limit

$$\lim_{\mu \downarrow 0} {}^{\mu}\iota(x) = \iota_{\mathcal{X}}(x)$$

for all $x \in E$.

5. *Local convexity and differentiability around points of interest.* Adding any quadratic regularizer to ${}^{\mu}\iota$ makes the sum locally convex and differentiable around points of interest. To be precise, if at x, the set $\mathcal{X}$ is prox-regular, then for any value of $\beta > 0$, the function ${}^{\mu}\iota(x) + \frac{\beta}{2}\|x\|^2$ is convex and differentiable on a neighborhood of x.

The favorable features of ${}^{\mu}\iota$ motivate us to consider the following penalization formulation of problem ($\mathcal{P}$) denoted by problem ($\mathcal{P}_\mu$), where the subscript μ indicates the penalty parameter:

$$\text{minimize } f(x) + {}^{\mu}\mathcal{I}(x), \tag{$\mathcal{P}_\mu$}$$

where ${}^{\mu}\mathcal{I} \equiv {}^{\mu}\iota + (\beta/2)\|\cdot\|^2$, $x \in \mathbf{E}$ is the decision variable, and μ is a positive *penalty parameter*. We call the cost function in problem ($\mathcal{P}_\mu$) an *exterior-point minimization function*; the term is inspired by [29, §4.1]. The notation ${}^{\mu}\mathcal{I} \equiv {}^{\mu}\iota + (\beta/2)\|\cdot\|^2$ introduced in problem ($\mathcal{P}_\mu$) not only reduces notational clutter, but also alludes to a specific way of splitting the objective into two summands f and ${}^{\mu}\mathcal{I}$, which will ultimately allow us to establish convergence of our algorithm in Sect. 3. Because ${}^{\mu}\iota$ is an asymptotically exact approximation of $\iota_{\mathcal{X}}$ as $\mu \to 0$, solving problem ($\mathcal{P}_\mu$) for a small enough value of the penalty parameter μ suffices for all practical purposes.

To provide intuition on how the exterior-point minimization function in problem ($\mathcal{P}_\mu$) compares against the original minimization function in problem ($\mathcal{P}$), we provide an illustrative one-dimensional example in Fig. 1. Figure 1 captures all the key properties of our penalization scheme. In this figure, $f = (1/2)(\cdot)^2$, $\beta = 1$,

$\mathcal{X} = [-2, -1] \bigcup [2, 3]$. The problem has two local minima, one at -1 and one at -2. We see that for larger values of μ, problem $(\mathcal{P}_\mu)$ is not a good approximation of problem $(\mathcal{P})$, but around each local minimum there is a relatively large region where $f + {}^{\mu}\mathcal{I}$ is strongly convex and smooth. As μ gets smaller, problem $(\mathcal{P}_\mu)$ becomes a more accurate approximation of problem $(\mathcal{P})$, though the regions of convexity and smoothness around local minima shrink. For $\mu = 10^{-4}$, problem $(\mathcal{P}_\mu)$ is identical to problem $(\mathcal{P})$ for all practical purposes. Note that the regularization parameter β is kept fixed at its initial value 1 throughout.

Now that we have intuitively justified intuition behind working with $(\mathcal{P}_\mu)$, we are in a position to present our algorithm.

given: regularization parameter $\beta > 0$, an initial point z_{init}, initial penalty parameter μ_{init}, minimum penalty parameter $\mu_{\min}$, tolerance for the fixed point gap ϵ for each inner iteration, tolerance for stopping criterion δ for the outer iteration, and multiplicative factor $\rho \in (0, 1)$.

Initialization. $\mu{:=}\mu_{\text{init}}$, *and* $z^0{:=}z_{\text{init}}$.

Outer iteration. **while** *stopping criterion is not met* **do**

  Inner iteration. Using Algorithm 2, compute x_μ, y_μ, and z_μ that solve problem $(\mathcal{P}_\mu)$ with tolerance ϵ, where $z_\mu^0{:=}z^0$ is input as the initial point.

  Stopping criterion. **quit** *if* $|\left(f(\mathbf{\Pi}_{\mathcal{X}}\, x_\mu) + (\beta/2)\|\mathbf{\Pi}_{\mathcal{X}} x_\mu\|^2\right) - \left(f(x_\mu) + {}^{\mu}\mathcal{I}(x_\mu)\right)| \le \delta$.

  Set initial point for next inner iteration. $z^0{:=}z_\mu$.

  Update μ. $\mu{:=}\rho\mu$.

end

return x_μ, y_μ, and z_μ

Algorithm 1: Nonconvex Exterior-point Optimization Solver (NExOS). Here $\mathbf{\Pi}_{\mathcal{X}}(x)$ denotes the Euclidean projection of x on the nonconvex set $\mathcal{X}$.

given: starting point z^0, tolerance for the fixed point gap ϵ, and proximal parameter $\gamma > 0$.

Initialization. $n{:=}0$, $\kappa{:=}1/(\beta\gamma + 1)$, $\theta{:=}\mu/(\gamma\kappa + \mu)$.

while $\|x^n - y^n\| > \epsilon$ **do**

  Compute $x^{n+1}{:=}\mathbf{prox}_{\gamma f}\left(z^n\right)$.

  Compute $\tilde{y}^{n+1}{:=}\kappa\left(2x^{n+1} - z^n\right)$.

  Compute $y^{n+1}{:=}\theta\tilde{y}^{n+1} + (1-\theta)\,\mathbf{\Pi}_{\mathcal{X}}\left(\tilde{y}^{n+1}\right)$.

  Compute $z^{n+1}{:=}z^n + y^{n+1} - x^{n+1}$.

  Update $n{:=}n+1$.

end

return x^n, y^n, and z^n.

Algorithm 2: Inner Algorithm for problem $(\mathcal{P}_\mu)$. Here $\mathbf{\Pi}_{\mathcal{X}}(x)$ denotes the Euclidean projection of x on the nonconvex set $\mathcal{X}$, and $\mathbf{prox}_{\gamma f}$ denotes the proximal operator of f with parameter $\gamma > 0$ as defined in (4).

Algorithm description Algorithm 1 outlines NExOS. The main part is an outer loop that solves a sequence of penalized problems of the form problem $(\mathcal{P}_\mu)$ with strictly

decreasing penalty parameter μ, until the termination criterion is met, at which point the exterior-point minimization function is a sufficiently close approximation of the original cost function. For each μ, problem ($\mathcal{P}_\mu$) is solved by an inner algorithm, denoted by Algorithm 2.

One can derive Algorithm 2 by applying Douglas–Rachford splitting (DRS) [3, page 401] to problem ($\mathcal{P}_\mu$) as follows. If we apply Douglas–Rachford splitting [3, page 401] to problem ($\mathcal{P}_\mu$) with penalty parameter μ, we have the following variant with three sub-iterations:

$$\begin{aligned} x^{n+1} &= \mathbf{prox}_{\gamma f}\left(z^n\right) \\ y^{n+1} &= \mathbf{prox}_{\gamma\, {}^{\mu}\mathcal{I}}\left(2x^{n+1} - z^n\right) \\ z^{n+1} &= z^n + y^{n+1} - x^{n+1}. \end{aligned} \tag{DRS}$$

The computational cost for $\mathbf{prox}_{\gamma\, {}^{\mu}\mathcal{I}}$ is the same as computing a projection onto the constraint set $\mathcal{X}$, as stated in Lemma 1 below; this result follows from [5, Theorem 6.13, Theorem 6.63]. It should be noted that [5, Theorem 6.13, Theorem 6.63] assume convexity of the functions in the theorem statements, but its proof does not require convexity and works for nonconvex functions as well.

Lemma 1 (Computing $\mathbf{prox}_{\gamma\, {}^{\mu}\mathcal{I}}(x)$) *Consider the nonconvex compact constraint set $\mathcal{X}$ in problem ($\mathcal{P}$). Denote $\kappa = 1/(\beta\gamma + 1) \in [0, 1]$ and $\theta = \mu/(\gamma\kappa + \mu) \in [0, 1]$. Then, for any $x \in \mathbf{E}$, and for any $\mu, \beta, \gamma > 0$, we have $\mathbf{prox}_{\gamma\, {}^{\mu}\mathcal{I}}(x) = \theta\kappa x + (1 - \theta)\, \mathbf{\Pi}_{\mathcal{X}}(\kappa x)$.*

Finally, combining (DRS), [5, Theorem 6.13], and Lemma 1, we arrive at Algorithm 2.

Algorithm subroutines The inner algorithm requires two subroutines, evaluating (i) $\mathbf{prox}_{\gamma f}(x)$, which is the proximal operator of the convex function f at the input point x, and (ii) $\mathbf{\Pi}_{\mathcal{X}}(x)$, which is a projection of x on the nonconvex set $\mathcal{X}$. We discuss now how we compute them in our implementation. To that goal, we recall that, for a function g (not necessarily convex) its proximal operator $\mathbf{prox}_{\gamma g}$ and Moreau envelope ${}^{\gamma}g$, where $\gamma > 0$, are defined as:

$$\begin{aligned} \mathbf{prox}_{\gamma g}(x) &= \operatorname*{argmin}_{y \in \mathbf{E}} \left(g(y) + (1/2\gamma)\|y - x\|^2 \right), \\ {}^{\gamma}g(x) &= \min_{y \in \mathbf{E}} \left(g(y) + (1/2\gamma)\|y - x\|^2 \right). \end{aligned} \tag{4}$$

Computing proximal operator of f For the convex function f, $\mathbf{prox}_{\gamma f}$ is always single-valued and computing it is equivalent to solving a convex optimization problem, which often can be done in closed form for many relevant cost functions in machine learning [5, pp. 449–450]. If the proximal operator of f does not admit a closed form solution, then we solve the corresponding convex optimization problem (4) to a high precision solution. For this purpose, we can select any convex optimization solver

supported by `MathOptInterface`, which is the abstraction layer for optimization solvers in `Julia`.

Computing projection onto $\mathcal{X}$ The notation $\mathbf{\Pi}_{\mathcal{X}}(x)$ denotes the *projection operator* of x onto the constraint set $\mathcal{X}$, defined as

$$\mathbf{\Pi}_{\mathcal{X}}(x) = \mathbf{prox}_{\gamma \iota_{\mathcal{X}}}(x) = \text{argmin}_{y \in \mathcal{X}}(\|y - x\|^2).$$

A list of nonconvex sets that are easy to project onto can be found in [25, §4], this includes nonconvex sets such as boolean vectors with fixed cardinality, vectors with bounded cardinality, quadratic sets, matrices with bounded singular values, matrices with bounded rank etc. If $\mathcal{X}$ is in this list, then we project onto $\mathcal{X}$ directly.

Now consider the case where the constraint set $\mathcal{X}$ decomposes as $\mathcal{X} = \mathcal{C} \bigcap \mathcal{N}$, where $\mathcal{N}$ is a nonconvex set with tractable projection and $\mathcal{C}$ is any compact convex set. In this setup, let $\iota_{\mathcal{C}}$ and $\iota_{\mathcal{N}}$ be the indicator functions of $\mathcal{C}$ and $\mathcal{N}$, respectively. Defining $\phi = f + \iota_{\mathcal{C}}$, we write problem ($\mathcal{P}$) as: $\min_{x \in \mathbf{E}} \phi(x) + (\beta/2)\|x\|^2 + \iota_{\mathcal{N}}(x)$.

For any convex function ϕ, its Moreau envelope ${}^{\nu}\phi$, for any $\nu > 0$, has the following three desirable features.

1. For every $x \in \mathbf{E}$ we have ${}^{\nu}\phi(x) \leq \phi(x)$ and ${}^{\nu}\phi(x) \to \phi(x)$ as $\nu \to 0$ [53, Theorem 1.25].
2. we have $x^\star \in \text{argmin}_{x \in \mathbf{E}} \phi(x)$ if and only if $x^\star \in \text{argmin}_{x \in \mathbf{E}} {}^{\nu}\phi(x)$ with the minimizer $x^\star$ satisfying $\phi(x^\star) = {}^{\nu}\phi(x^\star)$ [3, Corollary 17.5].
3. the Moreau envelope ${}^{\nu}\phi$ is convex, and smooth (i.e., it is differentiable and its gradient is Lipschitz continuous) everywhere irrespective of the differentiability or smoothness of the original function ϕ. The gradient is: ${}^{\nu}\phi(x) = \left(x - \mathbf{prox}_{\nu\phi}(x)\right)/\nu$, which is $(1/\nu)-$Lipschitz continuous [3, Proposition 12.29].

These properties make ${}^{\nu}\phi$ a smooth approximation of ϕ for a small enough ν. Hence, we work with the following approximation of the original problem: $\min_x {}^{\nu}\phi + (\beta/2)\|x\|^2 + \iota_{\mathcal{N}}(x)$, where we replace f with ${}^{\nu}\phi$ and $\iota_{\mathcal{X}}$ with $\iota_{\mathcal{N}}$ in Algorithms 1 and 2. The proximal operator of ${}^{\nu}\phi$ can be computed using $\mathbf{prox}_{\gamma\, {}^{\nu}\phi}(x) = x + (\gamma/(\gamma + \nu))(\mathbf{prox}_{(\gamma+\nu)\phi}(x) - x)$, where computing $\mathbf{prox}_{(\gamma+\nu)\phi}(x)$ corresponds to solving the following convex optimization problem $\text{argmin}_{y \in C}\, \phi(y) + 1/(2(\gamma + \nu))\|y - x\|^2$, which follows from [3, Proposition 24.8].

Remark 1 (*Reasons for choosing Douglas–Rachford splitting as the inner algorithm*) Problem ($\mathcal{P}_\mu$) involves minimizing the sum of two functions: a convex function f and a nonconvex function ${}^{\mu}\mathcal{I}$. As the objective is split into two parts in problem ($\mathcal{P}_\mu$), selecting any other two-operator splitting algorithm (e.g., forward-backward splitting [56, page 25], Chambolle–Pock algorithm [56, page 32], ADMM [48] etc.) can work as the inner algorithm in principle. However, in the context of our problem setup, Douglas–Rachford splitting might be the most suitable choice for the following reasons.

1. We have picked Douglas–Rachford splitting over ADMM, because Douglas–Rachford operates on the original nonconvex problem, whereas ADMM can be viewed as Douglas–Rachford splitting on the dual of the original nonconvex problem [65]. As strong duality usually does not hold when the primal problem is

nonconvex, it seems more intuitive to work with the nonconvex problem directly over its dual.

2. We favored Douglas–Rachford splitting over proximal gradient method, because even when the problem is convex, Douglas–Rachford splitting converges under more general conditions, whereas proximal gradient method require more restrictive conditions to converge [57, page 49]. Hence, we believe that Douglas–Rachford splitting represents the most natural choice for the inner algorithm over the proximal gradient method.
3. Douglas–Rachford splitting is favorably unique in contrast with other two-operator splitting methods, as Douglas–Rachford splitting is the only two-operator splitting method that satisfies the following properties simultaneously [55]: (i) it is constructed only with scalar multiplication, addition, and proximal operators, (ii) it computes proximal operators only once every iteration, (iii) it converges unconditionally for maximally monotone operators, and (iv) it does not increase the problem size.
   In Sect. 3, some of these desirable properties of Douglas–Rachford splitting are exploited to establish convergence. While other operator splitting algorithms may work to establish convergence as well, some of the unique features of Douglas–Rachford splitting will be lost [55].

3 Convergence Analysis

This section is organized as follows. We start with the definition of the key geometry property of sets involving sparse and low-rank optimization problems. Then we define the local minima of such problems, followed by the assumptions we use in our convergence analysis. We next discuss the convergence roadmap, where the first step involves showing that the exterior point minimization function is locally strongly convex and smooth around local minima, and the second step entails connecting the local minima with the underlying operator controlling NExOS. Then, we present the main result, which shows that, under mild regularity conditions, the inner algorithm of NExOS finds local minima for the penalized problems at a linear convergence rate, and as the penalty parameter goes to zero, the local minima of the penalized problems converge to a local minimum of the original problem. Furthermore, we show that, when those regularity conditions do not hold, the inner algorithm is still guaranteed to subsequentially converge to a first-order stationary point at the rate $o(1/\sqrt{k})$.

The key geometric property of sparse and low-rank constraint sets that we use in our convergence analysis is *prox-regularity at local minima*, i.e., having single-valued Euclidean projection around local minima [50]. Prox-regularity of a set at a point is defined as follows.

Definition 1 (*Prox-regular set* [50]) A nonempty closed set $\mathcal{S} \subseteq \mathbf{E}$ is prox-regular at a point $x \in \mathcal{S}$ if projection onto $\mathcal{S}$ is single-valued on a neighborhood of x. The set $\mathcal{S}$ is prox-regular if it is prox-regular at every point in the set.

If the constraint set $\mathcal{X}$ decomposes as $\mathcal{X} = \mathcal{C} \bigcap \mathcal{N}$, where $\mathcal{C}$ is a compact convex set, and $\mathcal{N}$ is prox-regular around local minima, then the feasible set $\mathcal{X}$ inherits the

prox-regularity property around local minima from the set $\mathcal{N}$ (see Lemma 4 in Sect. 3). The set $\mathcal{N}$ in (2) is a prox-regular set at any point $X \in \mathbf{R}^{m\times d}$ where $\mathbf{rank}(X) = r$ [44, Proposition 3.8]. One can show that $\mathcal{X}$ inherits the prox-regularity property at any X with $\mathbf{rank}(X) = r$ from the set $\mathcal{N}$; a formal proof is given in Lemma 4 in Appendix A.1. Similarly, $\mathcal{N}$ in (1) is prox-regular at any point x satisfying $\mathbf{card}(x) = k$ because we can write $\mathbf{card}(x) \leq k$ as a special case of the low-rank constraint by embedding the components of x in the diagonal entries of a matrix and then using the prox-regularity of low-rank constraint set.

In our convergence analysis, we use the prox-regularity property of sparse and low-rank optimization to establish our convergence results, hence NExOS can be applied to problems involving other constraint sets that are prox-regular at local minimal. Some other notable prox-regular sets are as follows. Closed convex sets are prox-regular everywhere [53, page 612]. Examples of well-known prox-regular sets that are not convex include sets involving bilinear constraints [4], weakly convex sets [69], proximally smooth sets [23], strongly amenable sets [53, page 612], and sets with Shapiro property [59]. Also, a nonconvex set defined by a system of finitely many inequality and equality constraints for which a basic constraint qualification holds is prox-regular [52, page 10].

We next provide the definition of local minimum of problem ($\mathcal{P}$). Recall that, according to our setup the set $\mathcal{X}$ is prox-regular at local minimum.

Definition 2 (*Local minimum of problem* ($\mathcal{P}$)) A point $\bar{x} \in \mathcal{X}$ is a local minimum of problem ($\mathcal{P}$) if the set $\mathcal{X}$ is prox-regular at $\bar{x}$, and there exists a closed ball with center $\bar{x}$ and radius r, denoted by $\overline{B}(\bar{x}; r)$ such that for all $y \in \mathcal{X} \cap \overline{B}(\bar{x}; r) \setminus \{\bar{x}\}$, we have $f(\bar{x}) + (\beta/2)\|\bar{x}\|^2 < f(y) + (\beta/2)\|y\|^2$.

In the definition above, the strict inequality is due to the strongly convex nature of the objective $f + (\beta/2)\|\cdot\|^2$ and follows from [1, Proposition 2.1] and [53, Theorem 6.12]. We now state and justify the assumptions used in our convergence analysis.

Assumption 1 (*Strong convexity and smoothness of* f) The function f in problem ($\mathcal{P}_\mu$) is α-strongly convex and L-smooth where $L > \alpha > 0$, i.e., $f - (\alpha/2)\|\cdot\|^2$ is convex and $f - (L/2)\|\cdot\|^2$ is concave.

Assumption 2 (*Problem* ($\mathcal{P}$) *is not trivial*) The unique solution to the unconstrained strongly convex problem $\min_x f(x) + (\beta/2)\|x\|^2$ does not lie in $\mathcal{X}$.

Assumption 1 corresponds to the function $f + (\beta/2)\|\cdot\|^2$ being $(\alpha + \beta)$-strongly convex and $(L + \beta)$-smooth. In our convergence analysis, $\beta > 0$ can be arbitrarily small, so it does not fall outside the setup described in Sect. 1. The L-smoothness in f is equivalent to its gradient ∇f being $L-$Lipschitz everywhere on $\mathbf{E}$ [3, Theorem 18.15]. In our convergence analysis, this assumption is required in establishing linear convergence of the inner algorithms of NExOS.

Assumption 2 imposes that a local minimum of problem ($\mathcal{P}$) is not the global minimum of its unconstrained convex relaxation, which does not incur any loss of generality. We can solve the unconstrained strongly convex optimization problem $\min_x f(x) + (\beta/2)\|x\|^2$ and check if the corresponding minimizer lies in $\mathcal{X}$; if that is the case, then that minimizer is also the global minimizer of problem ($\mathcal{P}$), and there

is no point in solving the nonconvex problem. This can be easily checked by solving an unconstrained convex optimization problem, so Assumption 2 does not cause any loss of generality.

To discuss our convergence roadmap, we introduce some standard operator theoretic notions as follows. A set-valued operator $\mathbb{A} : \mathbf{E} \rightrightarrows \mathbf{E}$ maps an element x in $\mathbf{E}$ to a set $\mathbb{A}(x)$ in $\mathbf{E}$; its domain is defined as $\mathbf{dom}\,\mathbb{A} = \{x \in \mathbf{E} \mid \mathbb{A}(x) \neq \emptyset\}$, its range is defined as $\mathbf{ran}\,\mathbb{A} = \bigcup_{x\in\mathbf{E}} \mathbb{A}(x)$, and it is completely characterized by its graph: $\mathbf{gra}\,\mathbb{A} = \{(u, x) \in \mathbf{E} \times \mathbf{E} \mid u \in \mathbb{A}(x)\}$. Furthermore, we define $\mathbf{fix}\,\mathbb{A} = \{x \in \mathbf{E} \mid x \in \mathbb{A}(x)\}$, and $\mathbf{zer}\,\mathbb{A} = \{x \in \mathbf{E} \mid 0 \in \mathbb{A}(x)\}$. For every x, addition of two operators $\mathbb{A}_1, \mathbb{A}_2 : \mathbf{E} \rightrightarrows \mathbf{E}$, denoted by $\mathbb{A}_1 + \mathbb{A}_2$, is defined as $(\mathbb{A}_1 + \mathbb{A}_2)(x) = \mathbb{A}_1(x) + \mathbb{A}_2(x)$, subtraction is defined analogously, and composition of these operators, denoted by $\mathbb{A}_1\mathbb{A}_2$, is defined as $\mathbb{A}_1\mathbb{A}_2(x) = \mathbb{A}_1(\mathbb{A}_2(x))$; note that order matters for composition. Also, if $\mathcal{S} \subseteq \mathbf{E}$ is a nonempty set, then $\mathbb{A}(\mathcal{S}) = \cup\{\mathbb{A}(x) \mid x \in \mathcal{S}\}$.

We next discuss our convergence roadmap. Convergence of NExOS is controlled by the DRS operator of problem $(\mathcal{P}_\mu)$:

$$\mathbb{T}_\mu = \mathbf{prox}_{\gamma\,{}^{\mu}\mathcal{I}}\left(2\mathbf{prox}_{\gamma f} - \mathbb{I}\right) + \mathbb{I} - \mathbf{prox}_{\gamma f}, \tag{5}$$

where $\mu > 0$, and $\mathbb{I}$ stands for the identity operator in $\mathbf{E}$, i.e., for any $x \in \mathbf{E}$, we have $\mathbb{I}(x) = x$. Using $\mathbb{T}_\mu$, the inner algorithm—Algorithm 2—can be written as

$$1z^{n+1} = \mathbb{T}_\mu\left(z^n\right) \tag{$\mathcal{A}_\mu$}$$

where μ is the penalty parameter and z^n is initialized at the fixed point from the previous inner algorithm.

To show the convergence of NExOS, we first show that for some $\mu_{\max} > 0$, for any $\mu \in (0, \mu_{\max}]$, the exterior point minimization function $f + {}^{\mu}\mathcal{I}$ is strongly convex and smooth on some open ball with center x and radius $r_{\max}$, denoted by $B(\bar{x}; r_{\max})$, where it will attain a unique local minimum x_μ. Then we show that for $\mu \in (0, \mu_{\max}]$, the operator $\mathbb{T}_\mu(x)$ will be contractive in x and Lipschitz continuous in μ, and connects its fixed point set $\mathbf{fix}\,\mathbb{T}_\mu$ with the local minima x_μ, via the relationship $x_\mu = \mathbf{prox}_{\gamma f}(\mathbf{fix}\,\mathbb{T}_\mu)$. In the main convergence result, we show that for a sequence of penalty parameters $\mathfrak{M} = \{\mu_1, \mu_2, \mu_3, \ldots, \mu_N\}$ and under proper initialization, if we apply NExOS to $\mathfrak{M}$, then for all $\mu_m \in \mathfrak{M}$,the inner algorithm will linearly converge to x_{μ_m}, and as $\mu_N \to 0$, we will have $x_{\mu_N} \to \bar{x}$. Finally, we show that, when the regularity conditions of the prior result do not hold, the inner algorithm is still guaranteed to subsequentially converge to a first-order stationary point (not necessarily a local minimum) at the rate $o(1/\sqrt{k})$.

We next present a proposition that shows that the exterior point minimization function in problem $(\mathcal{P}_\mu)$ will be locally strongly convex and smooth around local minima for our selection of penalty parameters, even though problem $(\mathcal{P})$ is nonconvex. Furthermore, as the penalty parameter goes to zero, the local minimum of problem $(\mathcal{P}_\mu)$ converges to the local minimum of the original problem $(\mathcal{P})$. So, under proper initialization, NExOS can solve the sequence of penalized problems $\{\mathcal{P}_\mu\}_{\mu\in(0,\mu_{\text{init}}]}$ similar to convex optimization problems; we will prove this in our main convergence result (Theorem 1).

Proposition 1 (Attainment of local minimum by $f + {}^{\mu}\mathcal{I}$) *Let Assumptions 1 and 2 hold for problem ($\mathcal{P}$), and let $\bar{x}$ be a local minimum to problem ($\mathcal{P}$). Then the following hold.*

(i) There exist $\mu_{\max} > 0$ and $r_{\max} > 0$ such that for any $\mu \in (0, \mu_{\max}]$, the exterior point minimization function $f + {}^{\mu}\mathcal{I}$ in problem ($\mathcal{P}_\mu$) is strongly convex and smooth in the open ball $B(\bar{x}; r_{\max})$ and will attain a unique local minimum x_μ in this ball.

(ii) As $\mu \to 0$, this local minimum x_μ will go to $\bar{x}$ in limit, i.e., $x_\mu \to \bar{x}$.

Proof See Appendix B.2. □

Because the exterior point minimization function is locally strongly convex and smooth, intuitively the DRS operator of problem ($\mathcal{P}_\mu$) would behave similar to that of a DRS operator of a composite convex optimization problem, but locally. When we minimize a sum of two convex functions where one of them is strongly convex and smooth, the corresponding DRS operator is contractive [32, Theorem 1]. So, we can expect that the DRS operator for problem ($\mathcal{P}_\mu$) would be locally contractive around a local minimum, which indeed turns out to be the case as proven in the next proposition. Furthermore, the next proposition shows that $\mathbb{T}_\mu(x)$ is locally Lipschitz continuous in the penalty parameter μ around a local minimum for fixed x. As $\mathbb{T}_\mu(x)$ is locally contractive in x and Lipschitz continuous in μ, it ensures that as we reduce the penalty parameter μ, the local minimum x_μ of problem ($\mathcal{P}_\mu$) found by NExOS does not change abruptly.

Proposition 2 (Characterization of $\mathbb{T}_\mu$) *Let Assumptions 1 and 2 hold for problem ($\mathcal{P}$), and let $\bar{x}$ be a local minimum to problem ($\mathcal{P}$). Then the following hold.*

(i) There exists a contraction factor $\kappa' \in (0, 1)$ such that for any $x_1, x_2 \in B(\bar{x}; r_{\max})$ and $\mu \in (0, \mu_{\max}]$, we have $\left\|\mathbb{T}_\mu(x_1) - \mathbb{T}_\mu(x_2)\right\| \leq \kappa' \|x_1 - x_2\|$.

(ii) For any $x \in B(\bar{x}; r_{\max})$, the operator $\mathbb{T}_\mu(x)$ is Lipschitz continuous in μ, i.e., there exists an $\ell > 0$ such that for any $\mu_1, \mu_2 \in (0, \mu_{\max}]$ and $x \in B(\bar{x}; r_{\max})$, we have $\left\|\mathbb{T}_{\mu_1}(x) - \mathbb{T}_{\mu_2}(x)\right\| \leq \ell \|\mu_1 - \mu_2\|$.

Proof See Appendix B.3. □

If the inner algorithm ($\mathcal{A}_\mu$) converges to a point z_μ, then z_μ would be a fixed point of the DRS operator $\mathbb{T}_\mu$. Establishing the convergence of NExOS necessitates connecting the local minimum x_μ of problem ($\mathcal{P}_\mu$) to the fixed point set of $\mathbb{T}_\mu$, which is achieved by the next proposition. Because our DRS operator locally behaves in a manner similar to the DRS operator of a convex optimization problem as shown by Proposition 2, it is natural to expect that the connection between x_μ and z_μ in our setup would be similar to that of a convex setup, but in a local sense. This indeed turns out to be the case as proven in the next proposition. The statement of this proposition is structurally similar to [3, Proposition 25.1(ii)] that establishes a similar relationship globally for a convex setup, whereas our result is established around the local minima of problem ($\mathcal{P}_\mu$).

Proposition 3 (Relationship between local minima of problem ($\mathcal{P}$) and $\boldsymbol{fix}\,\mathbb{T}_\mu$) *Let Assumptions 1 and 2 hold for problem ($\mathcal{P}$). Let $\bar{x}$ be a local minimum to problem ($\mathcal{P}$), and $\mu \in (0, \mu_{\max}]$. Then, $x_\mu = \operatorname{argmin}_{B(\bar{x}; r_{\max})} f(x) + {}^{\mu}\mathcal{I}(x) = \mathbf{prox}_{\gamma f}\left(\boldsymbol{fix}\,\mathbb{T}_\mu\right)$, where the sets $\boldsymbol{fix}\,\mathbb{T}_\mu$, and $\mathbf{prox}_{\gamma f}\left(\boldsymbol{fix}\,\mathbb{T}_\mu\right)$ are singletons over $B(\bar{x}; r_{\max})$.*

Proof See Appendix B.4. □

Before we present the main convergence result, we provide a helper lemma, which shows how the distances between x_μ, z_μ and $\bar{x}$ change as μ is varied in Algorithm 1. Additionally, this lemma provides the range for the proximal parameter γ. If $\mathcal{X}$ is a bounded set satisfying $\|x\| \leq D$ for all $x \in \mathcal{X}$, then term $\max_{x \in B(\bar{x}; r_{\max})} \|\nabla f(x)\|$ in this lemma can be replaced with $L \times D$.

Lemma 2 (Distance between local minima of problem ($\mathcal{P}$) *with local minima of problem* ($\mathcal{P}_\mu$)) *Let Assumptions 1 and 2 hold for problem* ($\mathcal{P}$)*, and let* $\bar{x}$ *be a local minimum to problem* ($\mathcal{P}$) *over* $B(\bar{x}; r_{\max})$*. Then the following hold.*

(i) For any $\mu \in (0, \mu_{\max}]$*, the unique local minimum* x_μ *of problem* ($\mathcal{P}_\mu$) *over* $B(\bar{x}; r_{\max})$ *satisfies* $\|x_\mu - \bar{x}\| < r_{\max}/\eta'$ *for some* $\eta' > 1$.

(ii) Let z_μ *be the unique fixed point of* $\mathbb{T}_\mu$ *over* $B(\bar{x}; r_{\max})$ *corresponding to* x_μ*. Then for any* $\mu \in (0, \mu_{\max}]$*, we have* $r_{\max} - \|x_\mu - \bar{x}\| > (\eta' - 1) r_{\max}/\eta'$ *and* $r_{\max} - \|z_\mu - \bar{x}\| > \psi$*, where* $\psi = (\eta' - 1) r_{\max}/\eta' - \gamma \max_{x \in B(\bar{x}; r_{\max})} \|\nabla f(x)\| > 0$ *with the proximal parameter* γ *taken to satisfy*

$$0 < \gamma < (\eta' - 1) r_{\max} / \left(\eta' \max_{x \in B(\bar{x}; r_{\max})} \|\nabla f(x)\| \right).$$

Furthermore, $\min_{\mu \in (0, \mu_{\max}]} \left\{ \left(r_{\max} - \|z_\mu - \bar{x}\| \right) - \psi \right\} > 0$.

Proof See Appendix B.5. □

We now present our main convergence results for NExOS. For convenience, we denote the n-th iterates of the inner algorithm of NExOS for penalty parameter μ by $\{x_\mu^n, y_\mu^n, z_\mu^n\}$. In the theorem, an ϵ-approximate fixed point $\widetilde{z}$ of $\mathbb{T}_\mu$ is defined by $\max\{\|\widetilde{z} - \mathbb{T}_\mu(\widetilde{z})\|, \|z_\mu - \widetilde{z}\|\} \leq \epsilon$, where z_μ is the unique fixed point of $\mathbb{T}_\mu$ over $B(\bar{x}; r_{\max})$. Furthermore, define:

$$\overline{\epsilon} := \min\{ \min_{\mu \in (0, \mu_{\max}]} ((r_{\max} - \|z_\mu - \bar{x}\|) - \psi)/2, (1 - \kappa')\psi \} > 0, \tag{6}$$

where $\kappa' \in (0, 1)$ is the contraction factor of $\mathbb{T}_\mu$ for any $\mu > 0$ (cf. Proposition 2) and the right-hand side is positive due to the third and fifth equations of Lemma 2(ii). Theorem 1 states that if we have a good initial point z_{init} for the first penalty parameter μ_{init}, then NExOS will construct a finite sequence of penalty parameters such that all the inner algorithms for these penalty parameters will linearly converge to the unique local minima of the corresponding inner problems.

Theorem 1 (Convergence result for *NExOS*) *Let Assumptions 1 and 2 hold for problem* ($\mathcal{P}$)*, and let* $\bar{x}$ *be a local minimum to problem* ($\mathcal{P}$)*. Suppose that the fixed-point tolerance* ϵ *for Algorithm 2 satisfies* $\epsilon \in [0, \overline{\epsilon})$*, where* $\overline{\epsilon}$ *is defined in* (6)*. The proximal parameter* γ *is selected to satisfy the fourth equation of Lemma 2(ii). In this setup, NExOS will construct a finite sequence of strictly decreasing penalty parameters* $\mathfrak{M} = \{\mu_1 := \mu_{\text{init}}, \mu_2 = \rho\mu_1, \mu_3 = \rho\mu_2, \ldots\}$*, with* $\mu_{\text{init}} \leq \mu_{\max}$ *and* $\rho \in (0, 1)$*, such that we have the following recursive convergence property.*

For any $\mu \in \mathcal{M}$, if an ϵ-approximate fixed point of $\mathbb{T}_\mu$ over $B(\bar{x}; r_{\max})$ is used to initialize the inner algorithm for penalty parameter $\rho\mu$, then the corresponding inner algorithm iterates $z^n_{\rho\mu}$ linearly converges to $z_{\rho\mu}$ that is the unique fixed point of $\mathbb{T}_{\rho\mu}$ over $B(\bar{x}, r_{\max})$, and the iterates $x^n_{\rho\mu}, y^n_{\rho\mu}$ linearly converge to $x_{\rho\mu} = \mathbf{prox}_{\gamma f}(z_{\rho\mu})$, which is the unique local minimum to $(\mathcal{P}_{\rho\mu})$ over $B(\bar{x}; r_{\max})$.

Proof See Appendix B.6. □

From Theorem 1, we see that an ϵ-approximate fixed point of $\mathbb{T}_{\rho\mu}$ over $B(\bar{x}; r_{\max})$ can be computed and then used to initialize the next inner algorithm for penalty parameter $\rho^2\mu$; this chain of logic makes each inner algorithm linearly converge to the corresponding locally optimal solution. Finally, for the convergence of the first inner algorithm we have the following result, which states that if the initial point z_{init} is not "too far away" from $B(\bar{x}; r_{\max})$, then the first inner algorithm of NExOS for penalty parameter μ_1 converges to a locally optimal solution of $(\mathcal{P}_{\mu_1})$.

Lemma 3 (Convergence of the first inner algorithm) *Let $\bar{x}$ be a local minimum to problem $(\mathcal{P})$, where Assumptions 1 and 2 hold. Let z_{init} be the chosen initial point for $\mu_1 := \mu_{\text{init}}$ such that $\overline{B}(z_{\mu_1}; \|z_{\text{init}} - z_{\mu_1}\|) \subseteq B(\bar{x}; r_{\max})$, where z_{μ_1} be the corresponding unique fixed point of $\mathbb{T}_{\mu_1}$. Then, $z^n_{\mu_1}$ linearly converges to z_{μ_1} and both $x^n_{\mu_1}$ and $y^n_{\mu_1}$ linearly converge to the unique local minimum x_{μ_1} of $(\mathcal{P}_{\mu_1})$ over $B(\bar{x}; r_{\max})$.*

Proof See Appendix B.7. □

We now discuss what can be said if the initial point z_{init} does not necessarily satisfy the conditions stated in Theorem 1 or Lemma 3. Unfortunately, in such a situation, we can only show subsequential convergence of the iterates.

Theorem 2 (Convergence result for *NExOS* for z_{init} that is far away from $B(\bar{x}; r_{max})$) *Suppose, the proximal parameter γ is selected to satisfy $0 < \gamma < 1/L$ and let z_{init} be the any arbitrarily chosen initial point that does not satisfy the conditions of Lemma 3. Then, in this setup, NExOS will construct a finite sequence of strictly decreasing penalty parameters $\mathfrak{M} = \{\mu_1 := \mu_{\text{init}}, \mu_2 = \rho\mu_1, \mu_3 = \rho\mu_2, \ldots\}$, and $\rho \in (0, 1)$, such that we have the following recursive convergence property. For any $\mu \in \mathcal{M}$, if an ϵ-approximate fixed point of $\mathbb{T}_\mu$ over $B(\bar{x}; r_{\max})$ is used to initialize the inner algorithm for penalty parameter $\rho\mu$, then the corresponding inner algorithm iterates $z^n_{\rho\mu}$ subsequentially converges to $z_{\rho\mu}$ that is a fixed point of $\mathbb{T}_{\rho\mu}$, and the iterates $x^n_{\rho\mu}, y^n_{\rho\mu}$ subsequentially converge to a first-order stationary point to $(\mathcal{P}_{\rho\mu})$ denoted by $x_{\rho\mu} = \mathbf{prox}_{\gamma f}(z_{\rho\mu})$ with the rate $\min_{n \le k} \|\nabla \left(f + {}^{\mu}\mathcal{I}\right)(x^n_{\rho\mu})\| \le \frac{1-\gamma L}{2L} o(1/\sqrt{k})$.*

Proof See Appendix B.8. □

4 Numerical Experiments

In this section, we apply NExOS to the following nonconvex optimization problems of substantial current interest for both synthetic and real-world datasets: sparse regression problem in Sect. 4.1, affine rank minimization problem in Sect. 4.2, and low-rank

factor analysis problem in Sect. 4.3. We illustrate that NExOS produces solutions that are either competitive or better in comparison with the other approaches on different performance measures. We have implemented NExOS in `NExOS.jl` solver, which is an open-source software package written in the `Julia` programming language. `NExOS.jl` can address any optimization problem of the form of problem ($\mathcal{P}$). The code and documentation are available online at: https://github.com/Shuvomoy/NExOS.jl.

In our numerical experiments, we present a comprehensive evaluation of NExOS, showing both statistical and optimization-theoretic evaluations. This dual approach is deliberate—while our primary contribution is in developing optimization methodology, the optimization problems considered in this section-such as sparse regression, affine rank minimization, matrix completion, and factor analysis-are deeply rooted in the fields of statistics and machine learning [7, 28, 36–38, 40]. Therefore, our numerical experiments are constructed not only to demonstrate NExOS efficiently computing local minima for nonconvex problems but also to highlight its ability to provide statistically robust solutions, which are also important in the application context. This dual capacity is of paramount importance for practical applications in statistics and machine learning, underlining the algorithm's versatility and effectiveness. By addressing these aspects, we aim to illustrate the broad applicability of NExOS across optimization-theoretic and applied statistical or learning domains. Here, we stress that while our optimization-theoretic evaluations are grounded in both theory and empirical experiments, the statistical evaluations of NExOS are based on empirical observations made in the context of the experiments conducted in this section.

To compute the proximal operator of a function f with closed form or easy-to-compute solution, `NExOS.jl` uses the open-source package ProximalOperators.jl [61]. When f is a constrained convex function (i.e., a convex function over some convex constraint set) with no closed form proximal map, `NExOS.jl` computes the proximal operator by using the open-source `Julia` package `JuMP` [27] and any of the commercial or open-source solver supported by it. The set $\mathcal{X}$ can be any prox-regular nonconvex set fitting our setup. Our implementation is readily extensible using `Julia` abstract types so that the user can add support for additional convex functions and prox-regular sets. The numerical study is executed on a MacBook Pro laptop with Apple M1 Max chip with 32 GB memory. The datasets considered in this section, unless specified otherwise, are available online at: http://tinyurl.com/NExOSDatasets.

In applying NExOS, we use the following values that we found to be the best performing based on our empirical observations. We take the starting value of μ to be 2, and reduce this value with a multiplicative factor of 0.5 during each iteration of the outer loop until the termination criterion is met. The value of the proximal parameter γ is chosen to be 10^{-3}. We initialize our iterates at $\mathbf{0}$. Maximum number of inner iterations for a fixed value of μ is taken to be 1000. The tolerance for the fixed point gap for each penalized problem is taken to be 10^{-4} and the tolerance for the termination criterion is taken to be 10^{-6}.

Value of β is taken to be 10^{-8} for the following reasons. In Sect. 3, we showed that the presence of $\beta > 0$, ensures that each penalized subproblem is locally strongly convex and smooth, having a unique local minimum. This, in turn, helps to establish linear convergence of the inner algorithm for each subproblem. We empirically

demonstrate in this section that the impact of the condition $\beta > 0$, despite being critical in the theoretical analysis of our algorithm, seems to only be marginal as it can be made to be as small as 10^{-8}. We use this extremely small value of β to stress-test NExOS empirically and show that even for such a small value of β, our algorithm still works well in practice. Here, we stress that the default value of $\beta = 10^{-8}$ used in our numerical experiments should be viewed as a mere heuristic that seems to empirically work for the numerical experiments that we considered in our paper. We leave a more methodical investigation of the smallest admissible values of β or the effect of completely omitting it to future work.

4.1 Sparse Regression

In (SR), we set $\mathcal{X} := \{x \mid \|x\|_\infty \leq \Gamma, \mathbf{card}(x) \leq k\}$, and $f(x) := \|Ax - b\|_2^2$. A projection onto $\mathcal{X}$ can be computed using the formula in [40, §2.2], whereas the proximal operator for f can be computed using the formula in [48, §6.1.1]. Now we are in a position to apply NExOS to this problem.

4.1.1 Synthetic Dataset: Comparison with Elastic Net and `Gurobi`

We compare the solution found by NExOS with the solutions found by elastic net (glmnet used for the implementation) and spatial branch-and-bound algorithm (Gurobi used for the implementation). Elastic net is a well-known method for computing an approximate solution to the regressor selection problem (SR), which empirically works extremely well in recovering support of the original signal. On the other hand, Gurobi's spatial branch-and-bound algorithm is guaranteed to compute a globally optimal solution to (SR). NExOS is guaranteed to provide a locally optimal solution under regularity conditions; so to investigate how close NExOS can get to the globally minimum value we consider a parallel implementation of NExOS running on multiple (20) worker processes, where each process runs NExOS with different random initialization, and we take the solution associated with the least objective value.

Elastic net

Elastic net is a well-known method for solving the regressor selection problem, that computes an approximate solution as follows. First, elastic net solves:

$$\text{minimize } \|Ax - b\|_2^2 + \lambda\|x\|_1 + (\beta/2)\|x\|_2^2, \tag{7}$$

where λ is a parameter that is related to the sparsity of the decision variable $x \in \mathbf{R}^d$. To solve (7), we have used the open-source `R` pacakge glmnet [31].

To compute λ corresponding to $\mathbf{card}(x) \leq k$ we follow the method proposed in [36, §3.4] and [20, Example 6.4]. We solve the problem (7) for different values of λ, and find the smallest value of λ for which $\mathbf{card}(x) \leq k$, and we consider the sparsity pattern of the corresponding solution $\tilde{x}$. Let the index set of zero elements of $\tilde{x}$ be $\mathcal{Z}$, where $\mathcal{Z}$ has $d - k$ elements. Then the elastic net solves:

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_2^2 + (\beta/2)\|x\|_2^2 \\ \text{subject to} & (\forall j \in \mathcal{Z}) \quad x_j = 0, \end{array} \tag{8}$$

where $x \in \mathbf{R}^d$ is the decision variable. Solving this problem corresponds to solving a positive semidefinite linear system, which we solve using the built-in `LinearAlgebra` package in `Julia`.

Spatial branch-and-bound algorithm

The problem (SR) can also be modeled equivalently as the following mixed integer quadratic optimization problem [13]:

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_2^2 + (\beta/2)\|x\|_2^2 \\ \text{subject to} & |x_i| \leq \Gamma y_i, \quad i = 1, \ldots, d \\ & \sum_{i=1}^d y_i \leq k, \quad x \in \mathbf{R}^d, \quad y \in \{0,1\}^d, \end{array}$$

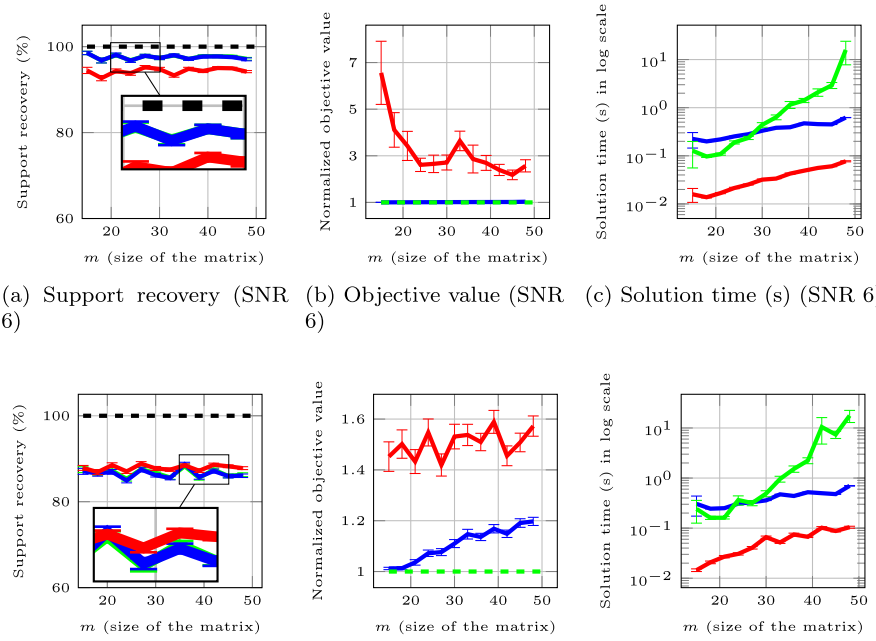
which can be solved to a certifiable global optimality using Gurobi's spatial branch-and-bound algorithm.

Data generation process and setup The data generation procedure is similar to [25, 37]. We consider two signal-to-noise ratio (SNR) regimes: SNR 1 and SNR 6, where for each SNR, we vary m from 25 to 50, and for each value of m, we generate 50 random problem instances. We limit the size of the problems because the solution time by Gurobi's spatial branch-and-bound algorithm becomes too large for comparison if we go beyond the aforementioned size. For a certain value of m, the matrix $A \in \mathbf{R}^{m \times 2m}$ is generated from an independent and identically distributed normal distribution with $\mathcal{N}(0, 1)$ entries. We choose $b = A\widetilde{x} + v$, where $\widetilde{x}$ is drawn uniformly from the set of vectors satisfying $\mathbf{card}(\widetilde{x}) \leq \lfloor m/5 \rceil$ and $\|\widetilde{x}\|_\infty \leq \Gamma$ with $\Gamma = 1$. The vector v corresponds to noise, and is drawn from the distribution $\mathcal{N}(0, \sigma^2 I)$, where $\sigma^2 = \|A\widetilde{x}\|_2^2/(\text{SNR}^2/m)$, which keeps the signal-to-noise ratio to approximately equal to SNR. We consider a parallel implementation of NExOS where we have 100 runs of NExOS distrubuted over 20 independent worker processes on 10 cores. Each run is initialized with a random initial points chosen from the uniform distribution over the interval $[-\Gamma, \Gamma]$. Gurobi's spatial branch-and-bound algorithm also uses 10 cores.

Results Figure 2 compares NExOS (shown in blue), glmnet (shown in red) and Gurobi (shown in green) for solving (SR). The results displayed in the figures are averaged over 50 simulations for each value of m, and also show one-standard-error bands that represent one standard deviation confidence interval around the mean.

Figures 2a and d show the support recovery (%) of the solutions found by NExOS, glmnet, and Gurobi for SNR 6 and SNR 1, respectively. Given a solution x and true signal x^{True}, the support recovery is defined as $\sum_{i=1}^d 1_{\{\text{sign}(x_i)=\text{sign}(x_i^{\text{True}})\}}/d$, where $1_{\{\cdot\}}$ evaluates to 1 if $(\cdot)$ is true and 0 else, and $\text{sign}(t)$ is 1 for $t > 0$, -1 for $t < 0$, and 0 for $t = 0$. So, higher the support recovery, better is the quality of the found solution. For both SNRs, NExOS and Gurobi have almost identical support recovery. For the high SNR, NExOS recovers most of the original signal's support and is better than glmnet consistently. On average, NExOS recovers 4% more of the support than glmnet. However, this behaviour changes for the low SNR, where glmnet recovers 1.26% more of the support than NExOS. This differing behavior in low and high SNR is consistent with the observations made in [37].

Figure 2b and e compare the quality of the solution found by the algorithms in terms of the normalized objective value (the objective value of the found solution divided



(a) Support recovery (SNR 6) (b) Objective value (SNR 6) (c) Solution time (s) (SNR 6)

(d) Support recovery (SNR 1) (e) Objective value (SNR 1) (f) Solution time (s) (SNR 1)

Fig. 2 Sparse regression problem: comparison between NExOS (shown in blue), glmnet (shown in red), and Gurobi (shown in green). The first and second rows correspond to SNR 6 and SNR 1, respectively. For each SNR, the first column compares support recovery, the second column shows how close the objective value of the solution found by each algorithm gets to the optimal objective value (normalized as 1), and the third column shows the solution time (s) of each algorithm

by the otimal objective value) for SNR 6 and SNR 1, respectively. As Gurobi's spatial branch-and-bound algorithm finds certifiably globally optimal solution to (SR), its normalized objective value is always 1, though the runtime is orders of magnitude slower than glmnet and NExOS (see the next paragraph). The closer the normalized objective value is to 1, better is the quality of the solution in terms of minimizing the objective value. We see that for the high SNR, on average NExOS is able to find a solution that is very close to the globally optimal solution, whereas the solution found by glmnet has worse objective value on average. For the low SNR, on average the normalized objective values of the solutions found by both NExOS and glmnet get worse, though NExOS does better than glmnet in this case as well.

Finally, in Fig. 2c and f, we compare the solution times (in seconds and on log scale) of the algorithms for SNR 6 and SNR 1, respectively. We see that glmnet is slightly faster than NExOS. This slower performance is due to the fact that NExOS is a general purpose method, whereas glmnet is specifically optimized for the convexified sparse regression problem with a specific cost function. For smaller problems, Gurobi is somewhat faster than NExOS, however once we go beyond $m \geq 27$, the solution time by Gurobi starts to increase drastically. Beyond $m \geq 50$, comparing the solution

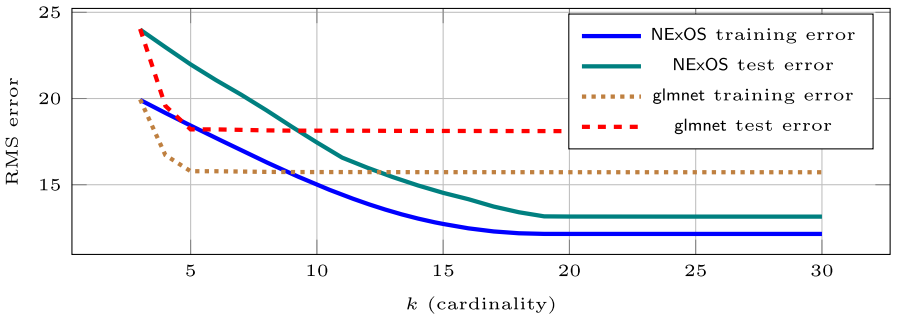


Fig. 3 RMS error vs k (cardinality) for the weather prediction problem

times is not meaningful as Gurobi cannot find a solution in 2 min, whereas NExOS takes less than 30 s.

4.1.2 Experiments and Results for Real-World Dataset

Description of the dataset We now investigate the performance of our algorithm on a real-world, publicly available dataset called the `weather prediction dataset`, where we consider the problem of predicting the temperature half a day in advance in 30 US and Canadian Cities along with 6 Israeli cities. The dataset contains hourly measurements of weather attributes *e.g.,* temperature, humidity, air pressure, wind speed, and so on. The dataset has $m = 45,231$ instances along with $d = 1,800$ attributes. The dataset is preprocessed in the same manner as described in [11, §8.3]. Our goal is to predict the temperature half a day in advance as a linear function of the attributes, where at most k attributes can be nonzero. We include a bias term in our model, i.e., in (SR) we set $A = [\bar{A} \mid \mathbf{1}]$. We randomly split 80% of the data into the training set and 20% of the data into the test set.

Results Figure 3 shows the RMS error for the training datasets and the test datasets for both NExOS and glmnet. The results for training and test datasets are reasonably similar for each value of k. This gives us confidence that the sparse regression model will have similar performance on new and unseen data. This also suggests that our model does not suffer from over-fitting. We also see that, for $k \geq 20$ and $k \geq 5$, none of the errors for NExOS and glmnet drop significantly, respectively. For smaller $k \leq 10$, glmnet does better than NExOS, but beyond $k \geq 10$, NExOS performs better than glmnet.

4.2 Affine Rank Minimization Problem

Problem description In (SR), we set $\mathcal{X} := \{X \in \mathbf{R}^{m\times d} \mid \mathbf{rank}(X) \leq r, \|X\|_2 \leq \Gamma\}$, and $f(X) := \|\mathcal{A}(X) - b\|_2^2$. To compute the proximal operator of f, we use the formula in [48, §6.1.1]. Finally, we use the formula in [25, page 14] for projecting onto $\mathcal{X}$. Now we are in a position to apply the NExOS to this problem.

Summary of the experiments performed First, we apply NExOS to solve (RM) for synthetic datasets, where we observe how the algorithm performs in recovering a low-rank matrix given noisy measurements and also compare NExOS with NCVX—an ADMM-based algorithm [25]. *Second,* we apply NExOS to a real-world dataset (`MovieLens 1M Dataset`) to see how our algorithm performs in solving a matrix-completion problem).

4.2.1 Experiments and Results for Synthetic Dataset

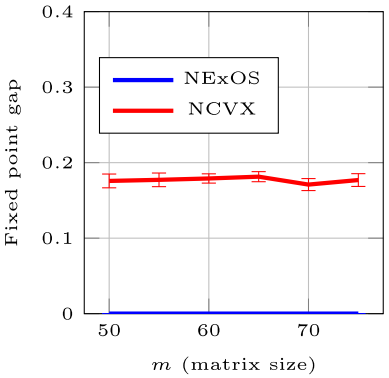
Data generation process and setup We generate the data as follows similar to [25]. We vary m (number of rows of the decision variable X) from 50 to 75 with a linear spacing of 5, where we take $d = 2m$, and rank to be equal to $m/10$ rounded to the nearest integer. For each value of m, we create 25 random instances as follows. The operator $\mathcal{A}$ is drawn from an iid normal distribution with $\mathcal{N}(0, 1)$ entries. Similarly, we create the low rank matrix X_{True} with rank r, first drawn from an iid normal distribution with $\mathcal{N}(0, 1)$ entries, and then truncating the singular values that exceed Γ to 0. Signal-to-noise ratio is taken to be around 20 by following the same method described for the sparse regression problem.

Results The results displayed in Fig. 4 average over 50 simulations for each value of m and also show one standard error band. We compare NExOS, with NCVX—an ADMM-based algorithm [25].

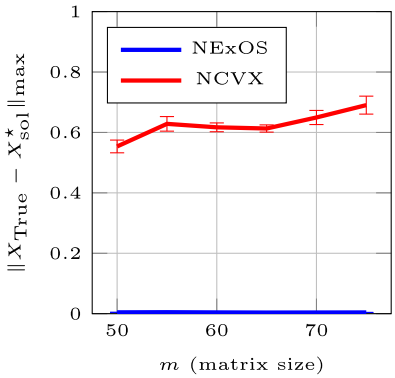
Figure 4a plots the normalized fixed point gap of the iterates for both algorithms computed by $\|X^\star_{\mathsf{Alg}} - Y^\star_{\mathsf{Alg}}\|/\|X_{\text{True}}\|$ with $\mathsf{Alg} \in \{\mathsf{NExOS}, \mathsf{NCVX}\}$ and $X^\star_{\mathsf{Alg}}, Y^\star_{\mathsf{Alg}}$ representing the final iterates produced by the algorithms. This plot shows that NCVX iterates have a fixed point gap larger than 0.17, i.e., the iterates do not converge within a reasonable fixed point gap. On the other hand, NExOS iterates converge with a normalized fixed-point gap reaching the desired tolerance of less than or equal to 10^{-4} for each instance.

Figure 4b shows how well NExOS and NCVX recovers the original matrix X_{True}. To quantify the recovery, we compute the max norm of the difference matrix $\|X_{\text{True}} - X^\star_{\mathsf{Alg}}\|_{\max} = \max_{i,j} |X_{\text{True}}(i, j) - X^\star_{\mathsf{Alg}}(i, j)|$, where the solution found by Alg is denoted by $X^\star_{\mathsf{Alg}}$. We see that the worst-case component-wise error is very small (smaller than 0.005 for each instance) in all the cases for NExOS, but for NCVX, it is larger than 0.5 for each instance. In other words, the solution found by NExOS is much closer to the ground truth as compared to NCVX.

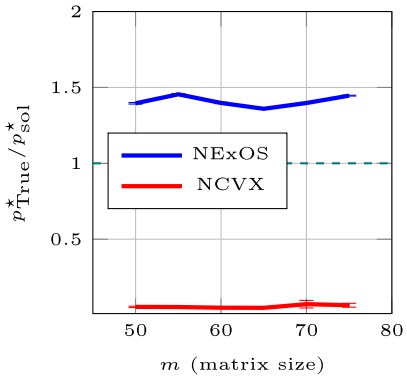
Finally, we show how the training loss of the solutions computed by NExOS and NCVX compare with the original matrix X_{True} in Fig. 4c. Note that for NExOS, the ratio $p^\star_{\text{True}}/p^\star_{\text{sol}}$ is larger than one in most cases, i.e., NExOS find a solutions with smaller cost compared to X_{True}. This is due to the fact that under the signal-to-noise ratio that we consider, the problem data can be explained better by another matrix with a lower training loss. That being said, $X^\star_{\mathsf{NExOS}}$ is not too far from X_{True} component-wise as we saw in Fig. 4b. On the other hand, for NCVX algorithm, the ratio $p^\star_{\text{True}}/p^\star_{\text{sol}}$ is smaller than 0.05 for each instance, i.e., the objective value of the solutions is 20 times worse than that of the original signal.



(a) Fixed point gap representing convergence of the iterates vs m



(b) Maximum absolute error in recovering the original matrix vs m



(c) Ratio of training losses of the true matrix X_{True} and the solution found by NExOS vs m

Fig. 4 Affine rank minimization problem: comparison between solutions found by NExOS and NCVX algorithm by [25]

4.2.2 Experiments and Results for Real-World Dataset: Matrix Completion Problem

Description of the dataset To investigate the performance of our problem on a real-world dataset, we consider the publicly available `MovieLens 1M Dataset`. This dataset contains 1,000,023 ratings for 3,706 unique movies; these recommendations were made by 6,040 MovieLens users. The rating is on a scale of 1 to 5. If we construct a matrix of movie ratings by the users (also called the preference matrix), denoted by Z, then it is a matrix of 6,040 rows (each row corresponds to a user) and 3,706 columns (each column corresponds to a movie) with only 4.47% of the total entries are observed,

while the rest being missing. Our goal is to complete this matrix, under the assumption that the matrix is low-rank. For more details about the model, see [40, §8.1].

To gain confidence in the generalization ability of this model, we use an out-of-sample validation process. By random selection, we split the available data into a training set (80% of the total data) and a test set (20% of the total data). We use the training set as the input data for solving the underlying optimization process, and the held-out test set is used to compute the test error for each value of r. The best rank r corresponds to the point beyond which the improvement is rather minor. We tested rank values r ranging in $\{1, 3, 5, 7, 10, 20, 25, 30, 35\}$. We compute the RMS error as follows. Let Ω_{test} be the index set corresponding to the test data. If $X^\star_{\mathsf{NExOS}}$ is the matrix returned by NExOS, then the corresponding RMS error is computed by using the formula

$$\text{RMS} = \sqrt{\frac{\sum_{(i,j)\in\Omega_{\text{test}}} \left(\left(X^\star_{\mathsf{NExOS}}\right)_{ij} - Z_{ij} \right)^2}{|\Omega_{\text{test}}|}},$$

where $|\Omega_{\text{test}}|$ is the number of elements in Ω_{test}.

Matrix completion problem The matrix completion problem is:

$$\begin{array}{l} \text{minimize } \sum_{(i,j)\in\Omega} (X_{ij} - Z_{ij})^2 + (\beta/2)\|X\|_F^2 \\ \text{subject to } \mathbf{rank}(X) \leq r, \quad \|X\|_2 \leq \Gamma, \end{array} \tag{MC}$$

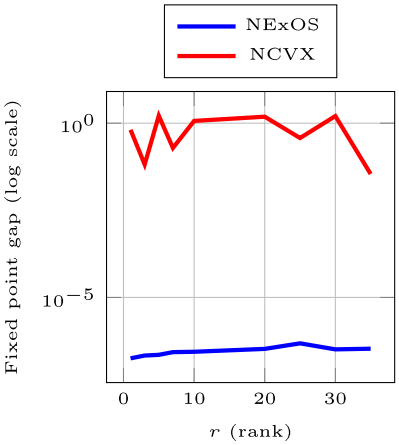
where $Z \in \mathbf{R}^{m\times d}$ is the matrix whose entries Z_{ij} are observable for $(i, j) \in \Omega$. Based on these observed entries, our goal is to construct a matrix $X \in \mathbf{R}^{m\times d}$ that has rank r. The problem above can be written as a special case of affine rank minimization problem (RM).

Results Figure 5 compares the solutions found by NExOS and NCVX.

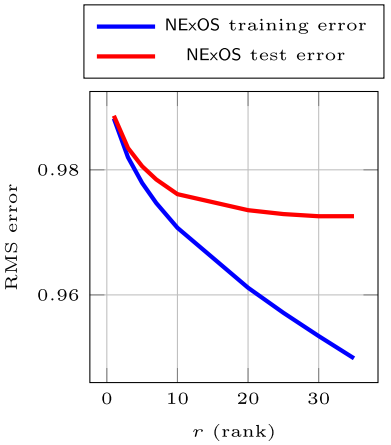
Figure 5a plots the normalized fixed point gap of the iterates for both algorithms calculated by $\|X^\star_{\mathsf{Alg}} - Y^\star_{\mathsf{Alg}}\|/\|X_{\text{True}}\|$ with $\mathsf{Alg} \in \{\mathsf{NExOS}, \mathsf{NCVX}\}$ and $X^\star_{\mathsf{Alg}}, Y^\star_{\mathsf{Alg}}$ representing the final iterates produced by the algorithms. This plot shows that NCVX iterates do not converge within a reasonable fixed point gap, whereas NExOS iterates converge for all the instances with a normalized fixed-point gap less than or equal to 10^{-6} for each instance.

Figure 5b shows the RMS error of NExOS for the training datatest and test dataset for each value of rank r. The results for training and test datasets are reasonably similar for each value of r. We observe that beyond rank 15, the reduction in the test error is rather minor and going beyond this rank provides only diminishing returns, which is a common occurrence for low-rank matrix approximation [42, §7.1]. Thus we can choose the optimal rank to be 15 for all practical purposes.

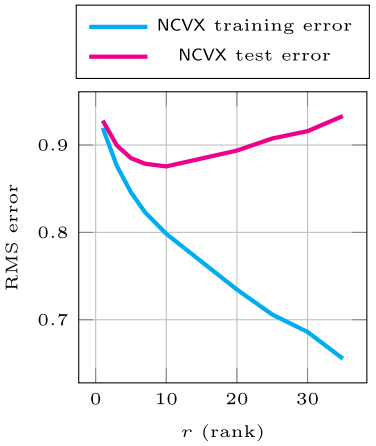
Figure 5c shows the RMS error of NCVX for the training dataset and test dataset for each value of rank r. We see that, unlike NExOS, the test error for NCVX keeps increasing with r, whereas the training error NCVX is smaller. Here we note that, because NCVX iterates do not reach a reasonable fixed point gap, the training or test error of NCVX may not provide meaningful information.



(a) Fixed point gap representing convergence of the iterates vs r



(b) Training and test error for NExOS vs r



(c) Training and test error for NCVX vs r

Fig. 5 Matrix completion problem: comparison between solutions found by NExOS and NCVX algorithm by [25]

4.3 Factor Analysis Problem

Problem description The factor analysis model with sparse noise (also known as low-rank factor analysis model) involves decomposing a given positive semidefinite matrix as a sum of a low-rank positive semidefinite matrix and a diagonal matrix with nonnegative entries [38, page 191]. It can be posed as [7]:

$$\begin{array}{ll}\text{minimize} & \|\Sigma - X - D\|_F^2 + (\beta/2)\left(\|X\|_F^2 + \|D\|_F^2\right) \\ \text{subject to} & D = \mathbf{diag}(d), \quad d \geq 0, \quad X \succeq 0, \quad \mathbf{rank}(X) \leq r \\ & \Sigma - D \succeq 0, \quad \|X\|_2 \leq \Gamma,\end{array} \tag{FA}$$

where $X \in \mathbf{S}^p$ and the diagonal matrix $D \in \mathbf{S}^p$ with nonnegative entries are the decision variables, and $\Sigma \in \mathbf{S}_+^p$, $r \in \mathbf{Z}_+$, and $\Gamma \in \mathbf{R}_{++}$ are the problem data. A proper solution for (FA) requires that both X and D are positive semidefinite. The term $\Sigma - D$ has to be positive semidefinite, else statistical interpretations of the solution is not impossible [64, page 326].

In (FA), we set $\mathcal{X} := \{(X, D) \in \mathbf{S}^p \times \mathbf{S}^p \mid \|X\|_2 \leq \Gamma, \mathbf{rank}(X) \leq r, D = \mathbf{diag}(d), d \geq 0\}$, and $f(X, D) := \|\Sigma - X - D\|_F^2 + I_{\mathcal{P}}(X, D)$, where $I_{\mathcal{P}}$ denotes the indicator function of the convex set $\mathcal{P} = \{(X, D) \in \mathbf{S}^p \times \mathbf{S}^p \mid X \succeq 0, D = \mathbf{diag}(d), d \geq 0, d \in \mathbf{R}^p\}$. To compute the projection onto $\mathcal{X}$, we use the formula in [25, page 14] and the fact that $\mathbf{\Pi}_{\{y|y\geq 0\}}(x) = \max\{x, 0\}$, where pointwise max is used. The proximal operator for f at (X, D) can be computed by solving:

$$\begin{array}{ll}\text{minimize} & \|\Sigma - \widetilde{X} - \widetilde{D}\|_F^2 + (1/2\gamma)\|\widetilde{X} - X\|_F^2 + (1/2\gamma)\|\widetilde{D} - D\|_F^2 \\ \text{subject to} & \widetilde{X} \succeq 0, \quad \widetilde{D} = \mathbf{diag}(\widetilde{d}), \quad \Sigma - \widetilde{D} \succeq 0, \quad \widetilde{d} \geq 0,\end{array}$$

where $\widetilde{X} \in \mathbf{S}_+^p$, and $\widetilde{d} \in \mathbf{R}_+^p$ (i.e., $\widetilde{D} = \mathbf{diag}(\widetilde{d})$) are the optimization variables. Now we are in a position to apply NExOS to this problem.

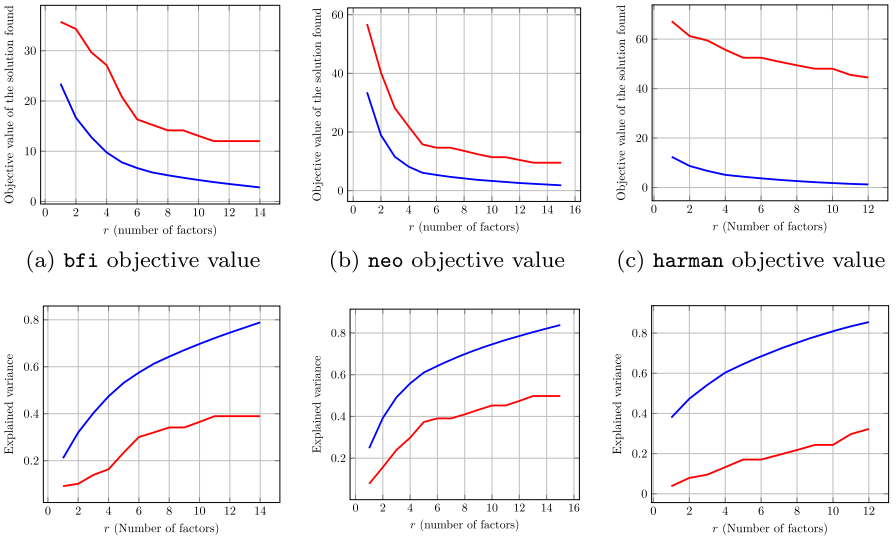
Comparison with nuclear norm heuristic We compare the solution provided by NExOS to that of the nuclear norm heuristic, which isthe most well-known heuristic to approximately solve (FA) [58] via following convex relaxation:

$$\begin{array}{ll}\text{minimize} & \|\Sigma - X - D\|_F^2 + \lambda \|X\|_* \\ \text{subject to} & D = \mathbf{diag}(d), \quad d \geq 0, \quad X \succeq 0, \\ & \Sigma - D \succeq 0, \quad \|X\|_2 \leq \Gamma,\end{array} \tag{9}$$

where λ is a positive parameter that is related to the rank of the decision variable X. Note that, as X is positive semidefinite, we have its nuclear norm $\|X\|_* = \mathbf{tr}(X)$.

Performance measures We consider two performance measures. First, we compare the training loss $\|\Sigma - X - D\|_F^2$ of the solutions found by NExOS and the nuclear norm heuristic. As both NExOS and the nuclear norm heuristic provide a point from the feasible set of (FA), such a comparison of training losses tells us which algorithm is providing a better quality solution. Second, we compute the *proportion of explained variance*, which represents how well the r-common factors explain the residual covariance, *i.e.*, $\Sigma - D$. For a given r, input proportion of variance explained by the r common factors is given by: $\sum_{i=1}^r \sigma_i(X) / \sum_{i=1}^p \sigma_i(\Sigma - D)$, where X, D are inputs, that correspond to solutions found by NExOS or the nuclear norm heuristic. As r increases, the explained variance increases to 1. The higher the value of the explained variance for a certain solution, the better is the quality of the solution.

Description of the datasets We consider three different real-world bench-mark datasets that are popularly used for factor analysis. The `bfi`, `neo` , and `Harman74` datasets



(a) `bfi` objective value (b) `neo` objective value (c) `harman` objective value

(d) `bfi` explained variance (e) `neo` explained variance (f) `harman` explained variance

Fig. 6 Figure showing performance of NExOS in solving factor analysis problem for different datasets. Each column represents one dataset. The first and second row compares training loss and proportion of the variance explained of the solutions found by NExOS (shown in blue) and the nuclear norm heuristic (shown in red)

contain (2800 observations, 28 variables), (1000 observations, 30 variables), and (145 observations, 24 variables), respectively.

Setup In applying NExOS for the factor analysis problem, we initialize our iterates with $Z_0 := \Sigma$ and $z_0 := \mathbf{0}$. All the other parameters are kept at their default values as stated in the beginning of Sect. 4. For each dataset, we vary the number of factors from 1 to $\lfloor p/2 \rfloor$, where p is the size of the underlying matrix Σ.

Results Figure 6 shows performance of NExOS in solving the factor analysis problem for different datasets, with each row representing one dataset. The first row compares the training loss of the solution found by NExOS and the nuclear norm heuristic. We see that for all the datasets, NExOS finds a solution with a training loss that is considerably smaller than that of the nuclear norm heuristic. The second row shows the proportion of variance explained by the algorithms considered for the datasets considered (higher is better). We see that in terms of the proportion of explained variance, NExOS delivers larger values than that of the nuclear norm heuristic for different values of r, which is indeed desirable. NExOS consistently provides solutions with better objective value and explained variance compared to the nuclear norm heuristic.

5 Conclusion

In this paper, we have presented NExOS, a first-order algorithm to solve optimization problems with convex cost functions over nonconvex constraint sets— a problem structure that is satisfied by a wide range of nonconvex optimization problems including sparse and low-rank optimization. We have shown that, under mild technical conditions, NExOS is able to find a locally optimal point of the original problem by solving a sequence of penalized problems with strictly decreasing penalty parameters. We have implemented our algorithm in the `Julia` package `NExOS.jl` and have extensively tested its performance on a wide variety of nonconvex optimization problems. We have demonstrated that NExOS is able to compute high quality solutions at a speed that is competitive with tailored algorithms.

A Proof and Derivation to Results in §1

A.1 Lemma Regarding Prox-Regularity of Intersection of Sets

Lemma 4 *Consider the nonempty constraint set $\mathcal{X} = \mathcal{C} \bigcap \mathcal{N} \subseteq \mathbf{E}$, where $\mathcal{C}$ is compact and convex, and $\mathcal{N}$ is prox-regular at $x \in \mathcal{X}$. Then $\mathcal{X}$ is prox-regular at x.*

Proof to Lemma 4 To prove this result we record the following result from [6], where by $d_{\mathcal{S}}(x)$ we denote the Euclidean distance of a point x from the set $\mathcal{S}$, and $\overline{\mathcal{S}}$ denotes closure of a set $\mathcal{X}$.

Lemma 5 (Intersection of Prox-Regular Sets [6, Corollary 7.3(a)]) *Let $\mathcal{S}_1, \mathcal{S}_2$ be two closed sets in $\mathbf{E}$, such that $\mathcal{S} = \mathcal{S}_1 \bigcap \mathcal{S}_2 \neq \emptyset$ and both $\mathcal{S}_1, \mathcal{S}_2$ are prox-regular at $x \in \mathcal{S}$. If $\mathcal{S}$ is metrically calm at x, i.e., if there exist some $\varsigma > 0$ and some neighborhood of x denoted by $\mathcal{B}$ such that $d_{\mathcal{S}}(y) \leq \varsigma(d_{\mathcal{S}_1}(y) + d_{\mathcal{S}_2}(y))$ for all $y \in \mathcal{B}$, then $\mathcal{S}$ is prox-regular at x.*

Proof to Lemma 4 By definition, projection onto $\mathcal{N}$ is single-valued on some open ball $B(x; a)$ with center x and radius $a > 0$ [50, Theorem 1.3]. The set $\mathcal{C}$ is compact and convex, hence projection onto $\mathcal{C}$ is single-valued around every point, hence single-valued on $B(x; a)$ as well [3, Theorem 3.14, Remark 3.15]. Note that for any $y \in B(x; a)$, $d_{\mathcal{X}}(y) = 0$ if and only if both $d_{\mathcal{C}}(y)$ and $d_{\mathcal{N}}(y)$ are zero. Hence, for any $y \in B(x; a) \bigcap \mathcal{X}$, the metrically calmness condition is trivially satisfied. Next, recalling that the distance from a closed set is continuous [53, Example 9.6], over the compact set $\overline{B(x; a) \setminus \mathcal{X}}$, define the function h, such that $h(y) = 1$ if $y \in \mathcal{X}$, and $h(y) = d_{\mathcal{X}}(y)/(d_{\mathcal{C}}(y) + d_{\mathcal{N}}(y))$ else. The function h is upper-semicontinuous over $\overline{B(x; a) \setminus \mathcal{X}}$, hence it will attain a maximum $\varsigma > 0$ over $\overline{B(x; a) \setminus \mathcal{X}}$ [54, Theorem 4.16], thus satisfying the metrically calmness condition on $B(x; a) \setminus \mathcal{X}$ as well. Hence, using Lemma 5, the constraint set $\mathcal{X}$ is prox-regular at x. □

B Proofs and Derivations to the Results in §3

B.1 Modifying NExOS for Nonsmooth and Convex Loss Function

We now discuss how to modify NExOS when the loss function is nonsmooth and convex. The key idea is working with a strongly convex, smooth, and arbitrarily close approximation of f; such smoothing techniques are very common in optimization [5, 47]. The optimization problem in this case, where the positive regularization parameter is denoted by $\widetilde{\beta}$, is given by: $\min_x \phi(x)+(\widetilde{\beta}/2)\|x\|^2+\iota_{\mathcal{X}}(x)$, where the setup is same as problem ($\mathcal{P}$), except the function $\phi : \mathbf{E} \to \mathbf{R}\cup\{+\infty\}$ is lower-semicontinuous, proper (its domain is nonempty), and convex. Let $\beta:=\widetilde{\beta}/2$. For a ν that is arbitrarily small, define the following β strongly convex and $(\nu^{-1}+\beta)$-smooth function: $f:={}^{\nu}\phi(\cdot)+(\beta/2)\|\cdot\|^2$ where ${}^{\nu}\phi$ is the Moreau envelope of ϕ with paramter ν. Following the properties of the Moreau envelope of a convex function discussed in §2, the following optimization problem acts as an arbitrarily close approximation to the first nonsmooth convex problem: $\min_x f+(\beta/2)\|x\|^2+\iota_{\mathcal{X}}(x)$, which has the same setup as problem ($\mathcal{P}$).

We can compute $\mathbf{prox}_{\gamma f}(x)$ using the formula in by [5, Theorem 6.13, Theorem 6.63]. Then, we apply NExOS to $\min_x f+(\beta/2)\|x\|^2+\iota_{\mathcal{X}}(x)$ and proceed in the same manner as discussed earlier.

B.2 Proof to Proposition 1

B.2.1 Proof to Proposition 1(i)

We prove (i) in three steps. In the *first step*, we show that for any $\mu > 0$, $f+{}^{\mu}\mathcal{I}$ will be differentiable on some $B(\bar{x}; r_{\text{diff}})$ with $r_{\text{diff}} > 0$. In the *second step*, we then show that, for any $\mu \in (0, 1/\beta]$, $f+{}^{\mu}\mathcal{I}$ will be strongly convex and differentiable on some $B(\bar{x}; r_{\text{cvxdiff}})$. In the *third step*, we will show that there exist $\mu_{\max} > 0$ such that for any $\mu \in (0, \mu_{\max}]$, $f+{}^{\mu}\mathcal{I}$ will be strongly convex and smooth on some $B(\bar{x}; r_{\max})$ and will attain the unique local minimum x_μ in this ball.

Proof of the first step

To prove the first step, we start with the following lemma regarding differentiability of ${}^{\mu}\iota$.

Lemma 6 (Differentiability of ${}^{\mu}\iota$) *Let $\bar{x}$ be a local minimum to problem ($\mathcal{P}$), where Assumptions 1 and 2 hold. Then there exists some $r_{\text{diff}} > 0$ such that for any $\mu > 0$: (i) the function ${}^{\mu}\iota$ is differentiable on $B(\bar{x}; r_{\text{diff}})$ with derivative $\nabla {}^{\mu}\iota = (1/\mu)(\mathbb{I}-\mathbf{\Pi}_{\mathcal{X}})$, and (ii) the projection operator $\mathbf{\Pi}_{\mathcal{X}}$ onto $\mathcal{X}$ is single-valued and Lipschitz continuous on $B(\bar{x}; r_{\text{diff}})$.*

Proof From [50, Theorem 1.3(e)], there exists some $r_{\text{diff}} > 0$ such that the function d^2 is differentiable on $B(\bar{x}; r_{\text{diff}})$. As ${}^{\mu}\iota = (1/2\mu)d^2$ from (3), it follows that for any $\mu > 0$, ${}^{\mu}\iota$ is differentiable on $B(\bar{x}; r_{\text{diff}})$ which proves the first part of (i). The second part of (i) follows from the fact that $\nabla d^2(x) = 2(x-\mathbf{\Pi}_{\mathcal{X}}(x))$ whenever d^2 is differentiable at x [50, page 5240]. Finally, from [50, Lemma 3.2], whenever d^2

is differentiable at a point, projection $\mathbf{\Pi}_{\mathcal{X}}$ is single-valued and Lipschitz continuous around that point, and this proves (ii). □

Due to the lemma above, $f + {}^{\mu}\mathcal{I}$ will be differentiable on $B(\bar{x}; r_{\text{diff}})$ with $r_{\text{diff}} > 0$, as f and $(\beta/2)\|\cdot\|^2$ are differentiable. Also, due to Lemma 6(ii), projection operator $\mathbf{\Pi}_{\mathcal{X}}$ is $\widetilde{L}$-Lipschitz continuous on $B(\bar{x}; r_{\text{diff}})$ for some $\widetilde{L} > 0$. This proves the first step.

Proof of the second step To prove this step, we are going to record: (1) the notion of general subdifferential of a function, followed by (2) the definition of prox-regularity of a function and its connection with prox-regular set, and (3) a helper lemma regarding convexity of the Moreau envelope under prox-regularity.

Definition 3 (*Fenchel, Fréchet, and general subdifferential*) For any lower-semicontinuous function $h : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$, its Fenchel subdifferential ∂h is defined as [24, page 1]: $u \in \partial h(x) \Leftrightarrow h(y) \geq h(x) + \langle u \mid y - x\rangle$ for all $y \in \mathbf{R}^n$. For the function h, its Fréchet subdifferential $\partial^F h$ (also known as regular subdifferential) at a point x is defined as [24, Definition 2.5]: $u \in \partial^F h(x) \Leftrightarrow \liminf_{y \to 0} (h(x+y) - h(x) - \langle u \mid y\rangle)/\|y\| \geq 0$. Finally, the general subdifferential of h, denoted by $\partial^G h$, is defined as [52, Equation (2.8)]: $u \in \partial^G h(x) \Leftrightarrow u_n \to u, x_n \to x, f(x_n) \to f(x)$, for some $(x_n, u_n) \in \mathbf{gra}\, \partial^F h$. If h is additionally convex, then $\partial h = \partial^F h = \partial^G h$ [24, Property (2.3), Property 2.6].

Definition 4 (*Connection between prox-regularity of a function and a set* [49, Definition 1.1]) A function $h : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ that is finite at $\tilde{x}$ is prox-regular at $\tilde{x}$ for $\tilde{\nu}$, where $\tilde{\nu} \in \partial^G h(\tilde{x})$, if h is locally l.s.c. at $\tilde{x}$ and there exist a distance $\sigma > 0$ and a parameter $\rho > 0$ such that whenever $\|x' - \tilde{x}\| < \sigma$ and $\|x - \tilde{x}\| < \sigma$ with $x' \neq x$, $\|h(x) - h(\tilde{x})\| < \sigma$, $\|\nu - \tilde{\nu}\| < \sigma$ with $\nu \in \partial^G h(x)$, we have $h(x') > h(x) + \langle \nu \mid x' - x\rangle - (\rho/2)\|x' - x\|^2$. Also, a set $\mathcal{S}$ is prox-regular at $\tilde{x}$ for $\tilde{\nu}$ if we have the indicator function $\iota_{\mathcal{S}}$ is prox-regular at $\tilde{x}$ for $\tilde{\nu} \in \partial^G \iota_{\mathcal{S}}(\tilde{x})$ [49, Proposition 2.11]. The set $\mathcal{S}$ is prox-regular at $\tilde{x}$ if it is prox-regular at $\tilde{x}$ for all $\tilde{\nu} \in \partial^G \iota_{\mathcal{S}}(\tilde{x})$ [53, page 612].

We have the following helper lemma from [49].

Lemma 7 ([49, Theorem 5.2]) *Consider a function h which is lower semicontinuous at 0 with $h(0) = 0$ and there exists $\rho > 0$ such that $h(x) > -(\rho/2)\|x\|^2$ for any $x \neq 0$. Let h be prox-regular at $\tilde{x} = 0$ and $\tilde{\nu} = 0$ with respect to σ and ρ (σ and ρ as described in Definition 4), and let $\lambda \in (0, 1/\rho)$. Then, on some neighborhood of 0, the function*

$$ {}^{\lambda}h + \rho/(2 - 2\lambda\rho)\|\cdot\|^2 \tag{10} $$

is convex, where ${}^{\lambda}h$ is the Moreau envelope of h with parameter λ.

Now we start proving step 2 earnestly. To prove this result, we assume $\bar{x} = 0$. This does not cause any loss of generality because this is equivalent to transferring the coordinate origin to the optimal solution and prox-regularity of a set and strong convexity of a function is invariant under such a coordinate transformation.

First, note that the indicator function of our constraint closed set $\mathcal{X}$ is lower semicontinuous due to [53, Remark after Theorem 1.6, page 11], and as $\bar{x}$, the local minimizer lies in $\mathcal{X}$, we have $\iota_{\mathcal{X}}(\bar{x}) = 0$. The set $\mathcal{X}$ is prox-regular at $\bar{x}$ for all $v \in \partial^G \iota_{\mathcal{X}}(x)$ per our setup, so using Definition 4, we have $\iota_{\mathcal{X}}$ prox-regular at $\bar{x} = 0$ for $\bar{v} = 0 \in \partial^G \iota_{\mathcal{X}}(\bar{x})$ (because $\bar{x} \in \mathcal{X}$, we will have 0 as a subgradient of $\partial \iota_{\mathcal{X}}(\bar{x})$) with respect to some distance $\sigma > 0$ and parameter $\rho > 0$.

Note that the indicator function satisfies $\iota_{\mathcal{X}}(x) = c\iota_{\mathcal{X}}(x)$ for any $c > 0$ due to its definition, so $u \in \partial^G \iota_{\mathcal{X}}(x) \Leftrightarrow cu \in c\partial^G \iota_{\mathcal{X}}(x) = \partial(c\iota^G_{\mathcal{X}}(x)) = \partial \iota^G_{\mathcal{X}}(x)$ [53, Equation 10(6)] In our setup, we have $\mathcal{X}$ prox-regular at $\bar{x}$. So, setting $h{:=}\iota_m athcal X$, $\tilde{x}{:=}\bar{x} = 0$, $\tilde{v}{:=}\bar{v} = 0$, and $v{:=}u/(\beta/2\rho)$ in Definition 4, we have $\iota_{\mathcal{X}}$ is also prox-regular at $\bar{x} = 0$ for $\bar{v} = 0$ with respect to distance $\sigma \min\{1, \beta/2\rho\}$ and parameter $\beta/2$.

Next, because the range of the indicator function is $\{0, \infty\}$, we have $\iota_{\mathcal{X}}(x) > -(\rho/2)\|x\|^2$ for any $x \neq 0$. So, we have all the conditions of Theorem 7 satisfied. Hence, applying Lemma 7, we have $(1/2\mu)\left(d^2 + \beta\mu/(2-\beta\mu)\|\cdot\|^2\right)$ convex and differentiable on

$$B\left(\bar{x}; \min\left\{\sigma \min\{1, \beta/2\rho\}, r_{\text{diff}}\right\}\right)$$

for any $\mu \in (0, 2/\beta)$, where r_{diff} comes from Lemma 6. As r_{diff} in this setup does not depend on μ, the ball does not depend on μ either. Finally, note that in our exterior-point minimization function we have ${}^{\mu}\mathcal{I} = (1/2\mu)\left(d^2 + \beta\mu\|\cdot\|^2\right)$.

So if we take $\mu \leq \frac{1}{\beta}$, then we have $(\beta/2)\mu/\left(1 - \mu(\beta/2)\right) \leq \beta\mu$, and on the ball $B\left(\bar{x}; \min\left\{\sigma \min\{1, \beta/2\rho\}, r_{\text{diff}}\right\}\right)$, the function ${}^{\mu}\mathcal{I}$ will be convex and differentiable. But f is strongly-convex and smooth, so $f + {}^{\mu}\mathcal{I}$ will be strongly convex and differentiable on $B\left(\bar{x}; \min\left\{\sigma \min\{1, \beta/2\rho\}, r_{\text{diff}}\right\}\right)$ for $\mu \in (0, 1/\beta]$. This proves step 2.

*Proof of the third step*As point $\bar{x} \in \mathcal{X}$ is a local minimum of problem ($\mathcal{P}$), from Definition 2, there is some $r > 0$ such that for all $y \in \overline{B}(\bar{x}; r)$, we have $f(\bar{x}) + (\beta/2)\|\bar{x}\|^2 < f(y) + (\beta/2)\|y\|^2 + \iota_{\mathcal{X}}(y)$.

Then, due to the first two steps, for any $\mu \in (0, 1/\beta]$, the function $f + {}^{\mu}\mathcal{I}$ will be strongly convex and differentiable on $B\left(\bar{x}; \min\left\{\sigma \min\{1, \beta/2\rho\}, r_{\text{diff}}\right\}\right)$. For notational convenience, denote $r_{\max} := \min\left\{\sigma \min\{1, \beta/2\rho\}, r_{\text{diff}}\right\}$, which is a constant. As $f + {}^{\mu}\mathcal{I}$ is a global underestimator of and approximates the function $f + (\beta/2)\|\cdot\|^2 + \iota_{\mathcal{X}}$ with arbitrary precision as $\mu \to 0$, the previous statement and [53, Theorem 1.25] imply that there exist some $0 < \mu_{\max} \leq 1/\beta$ such that for any $\mu \in (0, \mu_{\max}]$, the function $f + {}^{\mu}\mathcal{I}$ will achieve a local minimum x_μ over $B(\bar{x}; r_{\max})$ where $\nabla(f + {}^{\mu}\mathcal{I})$ vanishes, i.e.,

$$\nabla(f + {}^{\mu}\mathcal{I})(x_\mu) = \nabla f(x_\mu) + \beta x_\mu + (1/\mu)\left(x_\mu - \mathbf{\Pi}_{\mathcal{X}}\left(x_\mu\right)\right) = 0 \tag{11}$$

$$\Rightarrow x_\mu = (1/(\beta\mu + 1))\left(\mathbf{\Pi}_{\mathcal{X}}(x_\mu) - \mu\nabla f(x_\mu)\right). \tag{12}$$

As the right hand side of the last equation is a singleton, this minimum must be unique. Finally to show the smoothness $f + {}^{\mu}\mathcal{I}$, for any $x \in B(\bar{x}; r_{\max})$, we have

$$\nabla\left(f + {}^{\mu}\mathcal{I}\right)(x) \overset{a)}{=} \nabla f(x) + (\beta + (1/\mu))\, x - (1/\mu)\, \mathbf{\Pi}_{\mathcal{X}}(x), \tag{13}$$

where a) uses Lemma 6. Thus, for any $x_1, x_2 \in B(\bar{x}; r_{\max})$ we have $\|\nabla(f + (\beta/2)\| \cdot \|^2 + {}^{\mu}\iota)(x_1) - \nabla(f + (\beta/2)\| \cdot \|^2 + {}^{\mu}\iota)(x_2)\| \leq (L + \beta + (1/\mu) + \widetilde{L})\|x_1 - x_2\|$, where we have used the following: ∇f is L-Lipschitz everywhere due to f being an $L-$smooth function in $\mathbf{E}$ ([3, Theorem 18.15]), and $\mathbf{\Pi}_{\mathcal{X}}$ is $\widetilde{L}$-Lipschitz continuous on $B(\bar{x}; r_{\max})$, as shown in step 1. This completes the proof for (i).

(ii): Using [53, Theorem 1.25], as $\mu \to 0$, we have $x_\mu \to \bar{x}$, and $\left(f + {}^{\mu}\mathcal{I}\right)(x_\mu) \to f(\bar{x}) + (\beta/2)\|\bar{x}\|^2$. Note that x_μ reaches $\bar{x}$ only in limit, as otherwise Assumption 2 will be violated.

B.3 Proof to Proposition 2

B.3.1 Proof to Proposition 2(i)

We will need the notions of nonexpansive and firmly nonexpansive operators in this proof. An operator $\mathbb{A} : \mathbf{E} \to \mathbf{E}$ is nonexpansive on some set $\mathcal{S}$ if it is Lipschitz continuous with Lipschitz constant 1 on $\mathcal{S}$; the operator is contractive if the Lipschitz constant is strictly smaller than 1. On the other hand, $\mathbb{A}$ is firmly nonexpansive on $\mathcal{S}$ if and only if its reflection operator $2\mathbb{A} - \mathbb{I}$ is nonexpansive on $\mathcal{S}$. A firmly nonexpansive operator is always nonexpansive [3, page 59].

We next introduce the following definition.

Definition 5 (*Resolvent and reflected resolvent* [3, pages 333, 336]) For a lower-semicontinuous, proper, and convex function h, the resolvent and reflected resolvent of its subdifferential operator are defined by $\mathbb{J}_{\gamma\partial h} = (\mathbb{I} + \gamma\partial h)^{-1}$ and $\mathbb{R}_{\gamma\partial h} = 2\mathbb{J}_{\gamma\partial h} - \mathbb{I}$, respectively.

The proof of (i) is proven in two steps. First, we show that the reflection operator of $\mathbb{T}_\mu$, defined by

$$\mathbb{R}_\mu = 2\mathbb{T}_\mu - \mathbb{I}, \tag{14}$$

is contractive on $B(\bar{x}, r_{\max})$, and using this we show that $\mathbb{T}_\mu$ in also contractive there in the second step. To that goal, note that $\mathbb{R}_\mu$ can be represented as:

$$\mathbb{R}_\mu = (2\mathbf{prox}_{\gamma\,{}^{\mu}\mathcal{I}} - \mathbb{I})(2\mathbf{prox}_{\gamma f} - \mathbb{I}), \tag{15}$$

which can be proven by simply using (5) and (14) on the left-hand side and by expanding the factors on the right-hand side. Now, the operator $2\mathbf{prox}_{\gamma f} - \mathbb{I}$ associated with the α-strongly convex and L-smooth function f is a contraction mapping for any $\gamma > 0$ with the contraction factor $\kappa = \max\{(\gamma L - 1)/(\gamma L + 1), (1 - \gamma\alpha)/(\gamma\alpha + 1)\} \in (0, 1)$, which follows from [32, Theorem 1]. Next, we show that $2\mathbf{prox}_{\gamma\,{}^{\mu}\mathcal{I}} - \mathbb{I}$ is nonexpansive on $B(\bar{x}; r_{\max})$ for any $\mu \in (0, \mu_{\max}]$. For any $\mu \in (0, \mu_{\max}]$, define the function g as follows. We have $g(y) = {}^{\mu}\mathcal{I}(y)$ if $y \in B(\bar{x}; r_{\max}), g(y) = \liminf_{\tilde{y}\to y} {}^{\mu}\mathcal{I}(\tilde{y})$ if $\|y - \bar{x}\| = r_{\max}$, and $g(y) = \infty$ else. The function g is lower-semicontinuous, proper, and convex everywhere due to [3, Lemma 1.31 and Corollary 9.10]. As a result for $\mu \in (0, \mu_{\max}]$, we have $\mathbf{prox}_{\gamma g} = \mathbb{J}_{\gamma\partial g}$ on $\mathbf{E}$ and $\mathbf{prox}_{\gamma g}$ is

firmly nonexpansive and single-valued everywhere, which follows from [3, Proposition 12.27, Proposition 16.34, and Example 23.3]. But, for $y \in B(\bar{x}; r_{\max})$, we have ${}^{\mu}\mathcal{I}(y) = g(y)$ and $\nabla {}^{\mu}\mathcal{I}(y) = \partial g(y)$. Thus, on $B(\bar{x}; r_{\max})$, the operator $\mathbf{prox}_{\gamma\, {}^{\mu}\mathcal{I}} = \mathbb{J}_{\gamma \nabla {}^{\mu}\mathcal{I}}$, and it is firmly nonexpansive and single-valued for $\mu \in (0, \mu_{\max}]$. Any firmly nonexpansive operator $\mathbb{A}$ has a nonexpansive reflection operator $2\mathbb{A} - \mathbb{I}$ on its domain of firm nonexpansiveness [3, Proposition 4.2]. Hence, on $B(\bar{x}; r_{\max})$, for $\mu \in (0, \mu_{\max}]$ the operator $2\mathbf{prox}_{\gamma\, {}^{\mu}\mathcal{I}} - \mathbb{I}$ is nonexpansive using (15).

Now we show that $\mathbb{R}_\mu$ is contractive for every $x_1, x_2 \in B(\bar{x}; r_{\max})$ and $\mu \in (0, \mu_{\max}]$, we have $\|\mathbb{R}_\mu(x_1) - \mathbb{R}_\mu(x_2)\| \leq \|(2\mathbf{prox}_{\gamma f} - \mathbb{I})(x_1) - (2\mathbf{prox}_{\gamma f} - \mathbb{I})(x_2)\| \leq \kappa \|x_1 - x_2\|$ where the last inequality uses κ-contractiveness of $2\mathbf{prox}_{\gamma f} - \mathbb{I}$ thus proving that $\mathbb{R}_\mu$ acts as a contractive operator on $B(\bar{x}; r_{\max})$ for $\mu \in (0, \mu_{\max}]$. Similarly, for any $x_1, x_2 \in B(\bar{x}; r_{\max})$, using ($\mathcal{A}_\mu$) and the triangle inequality we have $\|\mathbb{T}_\mu(x_1) - \mathbb{T}_\mu(x_2)\| \leq (1 + \kappa)/2 \|x_1 - x_2\|$ and as $\kappa' = (1 + \kappa)/2 \in [0, 1)$; the operator $\mathbb{T}_\mu$ is κ'−contractive on on $B(\bar{x}; r_{\max})$, for $\mu \in (0, \mu_{\max}]$.

B.3.2 Proof to Proposition 2(ii)

Recalling $\mathbb{T}_\mu = (1/2)\mathbb{R}_\mu + (1/2)\mathbb{I}$ from (14), using (15), and then expanding, and finally using Lemma 1 and triangle inequality, we have for any $\mu, \tilde{\mu} \in (0, \mu_{\max}]$, $x \in B(\bar{x}; r_{\max})$, and $y = 2\mathbf{prox}_{\gamma f}(x) - x$:

$$\begin{aligned}\left\|\mathbb{T}_\mu(x) - \mathbb{T}_{\tilde{\mu}}(x)\right\| \leq\ & \|(\mu/(\gamma + \mu(\beta\gamma + 1)) - \tilde{\mu}/(\gamma + \tilde{\mu}(\beta\gamma + 1)))\| \, \|y\| \\ & + \|(\gamma/(\gamma + \mu(\beta\gamma + 1)) - \gamma/(\gamma + \tilde{\mu}(\beta\gamma + 1)))\| \, \|\mathbf{\Pi}_{\mathcal{X}}\,(y/(\beta\gamma + 1))\|\,. \end{aligned} \tag{16}$$

Now, in (16), the coefficient of $\|y\|$ satisfies $\|\mu/(\gamma + \mu(\beta\gamma + 1)) - \tilde{\mu}/(\gamma + \tilde{\mu}(\beta\gamma + 1))\| \leq (1/\gamma)\|\mu - \tilde{\mu}\|$

and similarly the coefficient of $\| \mathbf{\Pi}_{\mathcal{X}}\,(y/(\beta\gamma + 1)) \|$ satisfies

$$\|\gamma/(\gamma + \mu(\beta\gamma + 1)) - \gamma/(\gamma + \tilde{\mu}(\beta\gamma + 1))\| \leq (\beta + (1/\gamma))\|\mu - \tilde{\mu}\|.$$

Putting the last two inequalities in (16), and then replacing $y = 2\mathbf{prox}_{\gamma f}(x) - x$, we have for any $x \in \mathcal{B}$, and for any $\mu, \tilde{\mu} \in \mathbf{R}_{++}$,

$$\begin{aligned}&\left\|\mathbb{T}_\mu(x) - \mathbb{T}_{\tilde{\mu}}(x)\right\| \leq (1/\gamma)\, \|\mu - \tilde{\mu}\|\, \|y\| + (\beta + (1/\gamma))\, \|\mu - \tilde{\mu}\|\, \|\mathbf{\Pi}_{\mathcal{X}}\,(y/(\beta\gamma + 1))\| \\ &= \{(1/\gamma)\|2\mathbf{prox}_{\gamma f}(x) - x\| + (\beta + (1/\gamma))\|\, \mathbf{\Pi}_{\mathcal{X}}((2\mathbf{prox}_{\gamma f}(x) - x)/(\beta\gamma + 1))\|\}\|\mu - \tilde{\mu}\|. \end{aligned} \tag{17}$$

Now, as $B(\bar{x}; r_{\max})$ is a bounded set and $x \in \mathcal{B}$, norm of the vector $y = 2\mathbf{prox}_{\gamma f}(x) - x$ can be upper-bounded over $B(\bar{x}; r_{\max})$ because $2\mathbf{prox}_{\gamma f} - \mathbb{I}$ is continuous (in fact contractive) as shown in (i). Similarly, $\| \mathbf{\Pi}_{\mathcal{X}} \left((2\mathbf{prox}_{\gamma f}(x) - x)/(\beta\gamma + 1)\right) \|$ can be upper-bounded on $B(\bar{x}; r_{\max})$. Combining the last two-statements, it follows that there exists some $\ell > 0$ such that

$$\sup_{x\in B(\bar{x};r_{\max})} (1/\gamma)\|2\mathbf{prox}_{\gamma f}(x)-x\| + (\beta+1/\gamma)\left\|\mathbf{\Pi}_{\mathcal{X}}\left((2\mathbf{prox}_{\gamma f}(x)-x)/(\beta\gamma+1)\right)\right\| \leq \ell,$$

and putting the last inequality in (17), we arrive at the claim.

B.4 Proof to Proposition 3

The structure of the proof follows that of [3, Proposition 25.1(ii)]. Let $\mu \in (0, \mu_{\max}]$. Recalling Definition 5, and due to Proposition 1(i), $x_\mu \in B(\bar{x}; r_{\max})$ satisfies

$$\begin{aligned} x_\mu &= \operatorname*{argmin}_{B(\bar{x};r_{\max})} f(x) + {}^{\mu}\mathcal{I}(x) = \mathbf{zer}(\nabla f + \nabla\, {}^{\mu}\mathcal{I}) \\ &\overset{a)}{\Leftrightarrow} (\exists y \in \mathbf{E})\ \ x_\mu = \mathbb{J}_{\gamma\nabla\, {}^{\mu}\mathcal{I}}\mathbb{R}_{\gamma\nabla f}(y) \text{ and } x_\mu = \mathbb{J}_{\gamma\nabla f}(y), \end{aligned} \tag{18}$$

where $a)$ uses the facts (shown in the proof to Proposition 2) that: (i) $\mathbb{J}_{\gamma\nabla f}$ is a single-valued operator everywhere, whereas $\mathbb{J}_{\gamma\nabla\, {}^{\mu}\mathcal{I}}$ is a single-valued operator on the region of convexity $B(\bar{x}; r_{\max})$, and (ii) $x_\mu = \mathbb{J}_{\gamma\nabla f}(y)$ can be expressed as $x_\mu = \mathbb{J}_{\gamma\nabla f}(y) \Leftrightarrow 2x_\mu - y = \left(2\mathbb{J}_{\gamma\nabla f} - \mathbb{I}\right) y = \mathbb{R}_{\gamma\nabla f}(y)$. Also, using the last expression, we can write the first term of (18) as $\mathbb{J}_{\gamma\nabla\, {}^{\mu}\mathcal{I}}\mathbb{R}_{\gamma\nabla f}(y) = x_\mu \Leftrightarrow y \in \mathbf{fix}(\mathbb{R}_{\gamma\nabla\, {}^{\mu}\mathcal{I}}\mathbb{R}_{\gamma\nabla f})$. Because for lower-semicontinuous, proper, and convex function, the resolvent of the subdifferential is equal to its proximal operator [3, Proposition 12.27, Proposition 16.34, and Example 23.3], we have $\mathbb{J}_{\gamma\partial f} = \mathbf{prox}_{\gamma f}$ with both being single-valued. Using the last fact along with (18), $y \in \mathbf{fix}(\mathbb{R}_{\gamma\nabla\, {}^{\mu}\mathcal{I}}\mathbb{R}_{\gamma\nabla f})$, we have $x_\mu \in \mathbf{prox}_{\gamma f}\left(\mathbf{fix}\left(\mathbb{R}_{\gamma\nabla\, {}^{\mu}\mathcal{I}}\mathbb{R}_{\gamma\partial f}\right)\right)$, but x_μ is unique due to Proposition 1, so the inclusion can be replaced with equality. Thus x_μ, satisfies $x_\mu = \mathbf{prox}_{\gamma f}\left(\mathbf{fix}\left(\mathbb{R}_{\gamma\nabla\, {}^{\mu}\mathcal{I}}\mathbb{R}_{\gamma\partial f}\right)\right)$ where the sets are singletons due to Proposition 1 and single-valuedness of $\mathbf{prox}_{\gamma f}$. Also, because $\mathbb{T}_\mu$ in (5) and $\mathbb{R}_\mu$ in (14) have the same fixed point set (follows from (14)), using (15), we arrive at the claim.

B.5 Proof to Lemma 2

(i): This follows directly from the proof to Proposition 1.

(ii): From Lemma 2(i), and recalling that $\eta' > 1$, for any $\mu \in (0, \mu_{\max}]$, we have the first equation. Recalling Definition 5, and using the fact that for lower-semicontinuous, proper, and convex function, the resolvent of the subdifferential is equal to its proximal operator [3, Proposition 12.27, Proposition 16.34, and Example 23.3], we have $\mathbb{J}_{\gamma\partial f} = \mathbf{prox}_{\gamma f}$ with both being single-valued. So, from Proposition 3: $x_\mu = \mathbf{prox}_{\gamma f}(z_\mu) = (\mathbb{I} + \gamma\partial f)^{-1}(z_\mu) \Leftrightarrow z_\mu = x_\mu + \gamma\nabla f(x_\mu)$. Hence, for any $\mu \in (0, \mu_{\max}]$:

$$\begin{aligned} &\|z_\mu - \bar{x}\| = \|x_\mu + \gamma\nabla f(x_\mu) - \bar{x}\| \leq \|x_\mu - \bar{x}\| + \gamma\|\nabla f(x_\mu)\| \\ \Leftrightarrow &r_{\max} - \|z_\mu - \bar{x}\| \geq r_{\max} - \|x_\mu - \bar{x}\| - \gamma\|\nabla f(x_\mu)\| \overset{a)}{\geq} (\eta'-1)r_{\max}/\eta' - \gamma\|\nabla f(x_\mu)\|, \end{aligned}$$

where a) uses the first equation of Lemma 2(ii). Because, for the strongly convex and smooth function f, its gradient is bounded over a bounded set $B(\bar{x}; r_{\max})$ [51, Lemma 1, §1.4.2], then for γ satisfying the fourth equation of Lemma 2(ii) and the definition of ψ in the third equation of Lemma 2(ii), we have the second equation of Lemma 2(ii) for any $\mu \in (0, \mu_{\max}]$. To prove the final equation of Lemma 2(ii), note that

$$\begin{aligned}
&\lim_{\mu\to 0} \left(r_{\max} - \|z_\mu - \bar{x}\|\right) - \psi \\
&\overset{a)}{=} \lim_{\mu\to 0} \left(r_{\max} - \|x_\mu + \gamma\nabla f(x_\mu) - \bar{x}\|\right) - (\eta' - 1)r_{\max}/\eta' + \gamma \max_{x\in B(\bar{x};r_{\max})} \|\nabla f(x)\| \\
&\overset{b)}{=} (r_{\max} - \|\bar{x} + \gamma\nabla f(\bar{x}) - \bar{x}\|) - (\eta' - 1)r_{\max}/\eta' + \gamma \max_{x\in B(\bar{x};r_{\max})} \|\nabla f(x)\| \\
&= (1/\eta')r_{\max} + \gamma \left(\max_{x\in B(\bar{x};r_{\max})} \|\nabla f(x)\| - \|\nabla f(\bar{x})\|\right) > 0,
\end{aligned} \tag{19}$$

where in a) we have used $z_\mu = x_\mu + \gamma\nabla f(x_\mu)$ and the third equation of Lemma 2(ii), in b) we have used smoothness of f along with Proposition 1(ii). Inequality (19) along with the second equation of Lemma 2(ii) implies the final equation of Lemma 2(ii).

B.6 Proof to Theorem 1

We use the following result from [26] in proving Theorem 1.

Theorem 3 (Convergence of local contraction mapping [26, pp. 313–314]) *Let* $\mathbb{A} : \mathbf{E} \to \mathbf{E}$ *be some operator. If there exist* $\tilde{x}$, $\omega \in (0, 1)$, *and* $r > 0$ *such that (a)* $\mathbb{A}$ *is* ω*-contractive on* $B(\tilde{x}; r)$, *i.e., for all* x_1, x_2 *in* $B(\tilde{x}; r)$, *and (b)* $\|\mathbb{A}(\tilde{x}) - \tilde{x}\| \leq (1-\omega)r$. *Then* $\mathbb{A}$ *has a unique fixed point in* $B(\tilde{x}; r)$ *and the iteration scheme* $x_{n+1} = \mathbb{A}(x_n)$ *with the initialization* $x_0 := \tilde{x}$ *linearly converges to that unique fixed point.*

Furthermore, recall that NExOS (Algorithm 1) can be compactly represented using $(\mathcal{A}_\mu)$ as follows. For any $m \in \{1, 2, \ldots, N\}$ (equivalently for each $\mu_m \in \{\mu_1, \ldots, \mu_N\}$),

$$z_{\mu_m}^{n+1} = \mathbb{T}_{\mu_m}\left(z_{\mu_m}^{n}\right), \tag{20}$$

where $z_{\mu_m}^0$ is initialized at $z_{\mu_{m-1}}$. From Proposition 2, for any $\mu \in \mathfrak{M}$, the operator $\mathbb{T}_\mu$ is a κ'-contraction mapping over the region of convexity $B(\bar{x}; r_{\max})$, where $\kappa' \in (0, 1)$. From Proposition 1, there will be a unique local minimum x_μ of problem $(\mathcal{P}_\mu)$ over $B(\bar{x}; r_{\max})$. Suppose, instead of the exact fixed point $z_{\mu_{m-1}} \in \mathbf{fix}\,\mathbb{T}_{\mu_{m-1}}$, we have computed $\widetilde{z}$, which is an ϵ-approximate fixed point of $\mathbb{T}_{\mu_{m-1}}$ in $B(\bar{x}; r_{\max})$, i.e., $\|\widetilde{z} - \mathbb{T}_{\mu_{m-1}}(\widetilde{z})\| \leq \epsilon$ and $\|\widetilde{z} - z_{\mu_{m-1}}\| \leq \epsilon$, where $\epsilon \in [0, \bar{\epsilon})$. Then, we have:

$$\|\mathbb{T}_{\mu_{m-1}}(\widetilde{z}) - z_{\mu_{m-1}}\| = \|\mathbb{T}_{\mu_{m-1}}(\widetilde{z}) - \mathbb{T}_{\mu_{m-1}}(z_{\mu_{m-1}})\| \overset{a)}{\leq} \kappa' \underbrace{\|\widetilde{z} - z_{\mu_{m-1}}\|}_{\leq\epsilon} \leq \epsilon, \tag{21}$$

where $a)$ uses κ'-contractive nature of $\mathbb{T}_{\mu_{m-1}}$ over $B(\bar{x}; r_{\max})$. Hence, using triangle inequality,

$$\|\widetilde{z} - \bar{x}\| \overset{a)}{\leq} \|\widetilde{z} - \mathbb{T}_{\mu_{m-1}}(\widetilde{z})\| + \|\mathbb{T}_{\mu_{m-1}}(\widetilde{z}) - z_{\mu_{m-1}}\| + \|z_{\mu_{m-1}} - \bar{x}\| \overset{b)}{\leq} 2\epsilon + \|z_{\mu_{m-1}} - \bar{x}\|,$$

where $a)$ uses triangle inequality and $b)$ uses (21). As $\epsilon \in [0, \overline{\epsilon})$, where $\overline{\epsilon}$ is defined in (6), due to the second equation of Lemma 2(ii), we have $r_{\max} - \|\widetilde{z} - \bar{x}\| > \psi$.

Define $\Delta = \left((1-\kappa')\psi - \epsilon\right)/\ell$, which will be positive due to $\epsilon \in [0, \overline{\epsilon})$ and (6). Next, select $\theta \in (0, 1)$ such that $\overline{\Delta} = \theta\Delta < \mu_1$, hence there exists a $\rho \in (0, 1)$ such that $\overline{\Delta} = (1-\rho)\mu_1$. Now reduce the penalty parameter using

$$\mu_m = \mu_{m-1} - \rho^{m-2}\overline{\Delta} = \rho\mu_{m-1} = \rho^{m-1}\mu_1 \tag{22}$$

for any $m \geq 2$. Next, we initialize the iteration scheme $z^{n+1}_{\mu_m} = \mathbb{T}_{\mu_m}\left(z^n_{\mu_m}\right)$ at $z^0_{\mu_m} := \widetilde{z}$. Around this initial point, let us consider the open ball $B(\widetilde{z}, \psi)$. For any $x \in B(\widetilde{z}; \psi)$, we have $\|x - \bar{x}\| \leq \|x - \widetilde{z}\| + \|\widetilde{z} - \bar{x}\| < \psi + \|\widetilde{z} - \bar{x}\| < r_{\max}$, where the last inequality follows from $r_{\max} - \|\widetilde{z} - \bar{x}\| > \psi$. Thus we have shown that $B(\widetilde{z}; \psi) \subseteq B(\bar{x}; r_{\max})$. Hence, from Proposition 2, on $B(\widetilde{z}; \psi)$, the Douglas–Rachford operator $\mathbb{T}_{\mu_m}$ is contractive. Next, we have $\|\mathbb{T}_{\mu_m}(\widetilde{z}) - \widetilde{z}\| \leq (1-\kappa')\psi$, because $\|\mathbb{T}_{\mu_m}(\widetilde{z}) - \widetilde{z}\| \overset{a)}{\leq} \|\mathbb{T}_{\mu_m}(\widetilde{z}) - \mathbb{T}_{\mu_{m-1}}(\widetilde{z})\| + \|\mathbb{T}_{\mu_{m-1}}(\widetilde{z}) - \widetilde{z}\| \overset{b)}{\leq} \ell\|\mu_m - \mu_{m-1}\| + \epsilon \overset{c)}{\leq} \epsilon + \ell\Delta \overset{d)}{\leq} (1-\kappa')\psi$, where $a)$ triangle inequality, $b)$ uses Proposition 2(ii) and $\|\widetilde{z} - \mathbb{T}_{\mu_{m-1}}(\widetilde{z})\| \leq \epsilon$, $c)$ uses (22) and $\|\mu_m - \mu_{m-1}\| \leq \overline{\Delta} \leq \Delta$ $d)$ uses the definition of Δ. Thus, both conditions of Theorem 3 are satisfied, and $z^n_{\mu_m}$ in (20) will linearly converge to the unique fixed point z_{μ_m} of the operator $\mathbb{T}_{\mu_m}$, and $x^n_{\mu_m}$, $y^n_{\mu_m}$ will linearly converge to x_{μ_m}. This completes the proof.

B.7 Proof to Lemma 3

First, we show that, for the given initialization of z_{init}, the iterates $z^n_{\mu_1}$ stay in $\overline{B}(z_{\mu_1}; \|z_{\text{init}} - z_{\mu_1}\|)$ for any $n \in \mathbf{N}$ via induction. The base case is true via given. Let, $z^n_{\mu_1} \in \overline{B}(z_{\mu_1}; \|z_{\text{init}} - z_{\mu_1}\|)$. Then, $\|z^{n+1}_{\mu_1} - z_{\mu_1}\| \overset{a)}{=} \|\mathbb{T}_{\mu_1}(z^n_{\mu_1}) - \mathbb{T}_{\mu_1}(z_{\mu_1})\| \overset{b)}{\leq} \kappa'\|z^n_{\mu_1} - z_{\mu_1}\| \overset{c)}{\leq} \kappa'\|z_{\text{init}} - z_{\mu_1}\|$, where $a)$ uses $z_{\mu_1} \in \mathbf{fix}\,\mathbb{T}_\mu$, and $b)$ uses Proposition 2, and $c)$ uses $\|z^n_{\mu_1} - z_{\mu_1}\| \leq \|z_{\text{init}} - z_{\mu_1}\|$. So, the iterates $z^n_{\mu_1}$ stay in $\overline{B}(z_{\mu_1}; \|z_{\text{init}} - z_{\mu_1}\|)$. As, $\kappa' \in (0, 1)$, this inequality also implies that z^n_μ linearly converges to z_μ with the rate of at least κ'. Then using similar reasoning presented in the proof to Theorem 1, we have x^n_μ and y^n_μ linearly converge to the unique local minimum x_μ of problem $(\mathcal{P}_\mu)$. This completes the proof.

B.8 Proof to Theorem 2

The proof is based on the results in [43, Theorem 4] and [65, Theorem 4.3]. The function f is L-Lipschitz continuous and strongly smooth, hence f is a coercive function

satisfying $\liminf_{\|x\|\to\infty} f(x) = \infty$ and is bounded below [3, Corollary 11.17]. Also, ${}^{\mu}\mathcal{I}(x)$ is jointly continuous hence lower-semicontinuous in x and μ and is bounded below by definition. Let the proximal parameter γ be smaller than or equal to $1/L$. Then due to [43, (14), (15) and Theorem 4], $\{x_\mu^n, y_\mu^n, z_\mu^n\}$ (iterates of the inner algorithm of NExOS for any penalty parameter μ) will be bounded. This boundedness implies the existence of a cluster point of the sequence, which allows us to use [43, Theorem 4 and Theorem 1] to show that for any z_{init}, the iterates x_μ^n and y_μ^n subsequentially converges to a first-order stationary point x_μ satisfying $\nabla\left(f + {}^{\mu}\mathcal{I}\right)(x_\mu) = 0$. The rate $\min_{n\le k} \|\nabla\left(f + {}^{\mu}\mathcal{I}\right)(x_{\rho\mu}^n)\| \le ((1-\gamma L)/2\,L)o(1/\sqrt{k})$ is a direct application of [65, Theorem 4.3] as our setup satisfies all the conditions to apply it.

References

1. Auslender, A.: Stability in mathematical programming with nondifferentiable data. SIAM J. Control. Optim. **22**(2), 239–254 (1984)
2. Bach, F.: Sharp analysis of low-rank kernel matrix approximations. J. Mach. Learn. Res., (2013)
3. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, vol. 408. Springer, Berlin (2017)
4. Bauschke, H.H, Lal, M.K., Wang, X.: Projections onto hyperbolas or bilinear constraint sets in Hilbert spaces. J. Glob. Optim., pp. 1–12 (2022)
5. Beck, A.: First-Order Methods in Optimization, vol. 25. SIAM, Philadelphia (2017)
6. Bernard, F., Thibault, L., Zlateva, N.: Prox-regular sets and epigraphs in uniformly convex Banach spaces: various regularities and other properties. Trans. Am. Math. Soc. **363**(4), 2211–2247 (2011)
7. Bertsimas, D., Copenhaver, M.S., Mazumder, R.: Certifiably optimal low rank factor analysis. J. Mach. Learn. Res. **18**(1), 907–959 (2017)
8. Bertsimas, D., Cory-Wright, R.: A scalable algorithm for sparse portfolio selection. INFORMS J. Comput. **34**(3), 1489–1511 (2022)
9. Bertsimas, Dimitris, Cory-Wright, Ryan, Lo, Sean, Pauphilet, Jean: Optimal low-rank matrix completion: Semidefinite relaxations and eigenvector disjunctions. arXiv preprint arXiv:2305.12292, (2023)
10. Bertsimas, D., Cory-Wright, R., Pauphilet, J.: Mixed-projection conic optimization: A new paradigm for modeling rank constraints. Oper. Res. **70**(6), 3321–3344 (2022)
11. Bertsimas, D., Digalakis Jr, V., Li, M.L., Lami, O.S.: Slowly varying regression under sparsity. Oper. Res. (2024)
12. Bertsimas, D., Dunn, J.: Machine Learning Under a Modern Optimization Lens. Dynamic Ideas, Charlestown (2019)
13. Bertsimas, D., King, A., Mazumder, R.: Best subset selection via a modern optimization lens. Ann. Stat. 813–852 (2016)
14. Bertsimas, D., Parys, B.V.: Sparse hierarchical regression with polynomials. Mach. Learn., (2020)
15. Bertsimas, D., Van Parys, B., et al.: Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. Ann. Stat. **48**(1), 300–323 (2020)
16. Blanchard, J.D., Tanner, J., Wei, K.: CGIHT: Conjugate gradient iterative hard thresholding for compressed sensing and matrix completion. Inf. Inference **4**(4), 289–327 (2015)
17. Blumensath, T., Davies, M.E.: Iterative thresholding for sparse approximations. J. Fourier Anal. Appl. **14**(5–6), 629–654 (2008)
18. Blumensath, T., Davies, M.E.: Normalized iterative hard thresholding: Guaranteed stability and performance. IEEE J. Sel. Top. Signal Process. **4**(2), 298–309 (2010)
19. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends® Mach. Learn. 3(1):1–122 (2011)
20. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
21. Candès, E., Recht, B.: Exact matrix completion via convex optimization. *Found. Comput. Math.* (2009)
22. Candès, E., Wakin, M.B., Boyd, S.: Enhancing sparsity by reweighted l1 minimization. J. Fourier Anal. Appl. **14**, 877–905 (2008)

23. Clarke, F.H., Stern, R.J., Wolenski, P.R.: Proximal smoothness and the lower-$\mathcal{C}^2$ property. J. Convex Anal. **2**(1–2), 117–144 (1995)
24. Correa, R., Jofre, A., Thibault, L.: Characterization of lower semicontinuous convex functions. Proc. Am. Math. Soc. **116**, 67–72 (1992)
25. Diamond, S., Takapoui, R., Boyd, S.: A general system for heuristic minimization of convex functions over non-convex sets. Optim. Methods Softw. **33**(1), 165–193 (2018)
26. Dontchev, A.L., Rockafellar, R.T.: Implicit Functions and Solution Mappings, vol. 543. Springer, Berlin (2009)
27. Dunning, I., Huchette, J., Lubin, M.: JuMP: A modeling language for mathematical optimization. SIAM Rev. **59**(2), 295–320 (2017)
28. Fazel, M., Candes, E., Recht, B., Parrilo, P.: Compressed sensing and robust recovery of low rank matrices. In: 2008 42nd Asilomar Conference on Signals, Systems and Computers, 1043–1047 (2008)
29. Fiacco, A.V., McCormick, G.P.: Nonlinear Programming: Sequential Unconstrained Minimization Techniques. SIAM, Philadelphia (1990)
30. Foucart, S.: Hard thresholding pursuit: an algorithm for compressive sensing. SIAM J. Numer. Anal. **49**(6), 2543–2563 (2011)
31. Friedman, J., Hastie, T., Tibshirani, R., et al.: glmnet: Lasso and elastic-net regularized generalized linear models. R Package Version **1**(4), 1–24 (2009)
32. Giselsson, P., Boyd, S.: Linear convergence and metric selection for Douglas-Rachford splitting and ADMM. IEEE Trans. Autom. Control **62**(2), 532–544 (2017)
33. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
34. Gress, A., Davidson, I.: A flexible framework for projecting heterogeneous data. In: CIKM 2014 - Proceedings of the 2014 ACM International Conference on Information and Knowledge Management (2014)
35. Hardt, M., Meka, R., Raghavendra, P., Weitz, B.: Computational limits for matrix completion. J. Mach. Learn. Res. (2014)
36. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY (2001)
37. Hastie, T., Tibshirani, R., Tibshirani, R.J.: Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692 (2017)
38. Hastie, T., Tibshirani, R., Martin, W.: The Lasso and Generalizations. Statistical Learning with Sparsity. Taylor & Francis, New York (2015)
39. Hazimeh, H., Mazumder, R.: Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. Oper. Res. **68**(5), 1517–1537 (2020)
40. Jain, P., Kar, P.: Non-convex optimization for machine learning. Found. Trends® Mach. Learn. **10**(3–4), 142–336 (2017)
41. Jun, K.-S., Willett, R., Wright, S., Nowak, R.: Bilinear bandits with low-rank structure. In: International Conference on Machine Learning, pp. 3163–3172. PMLR (2019)
42. Lee, J., Kim, S., Lebanon, G., Singer, Y., Bengio, S.: LLORMA: Local low-rank matrix approximation. J. Mach. Learn. Res. **17**(1), 442–465 (2016)
43. Guoyin, L., Pong, T.K.: Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. Math. Programm. **159**(1), 371–401 (2016)
44. Russell Luke, D.: Prox-regularity of rank constraint sets and implications for algorithms. J. Math. Imaging Vis. **47**(3), 231–238 (2013)
45. Mazumder, R., Hastie, T., Tibshirani, R.: Spectral regularization algorithms for learning large incomplete matrices. J. Mach. Learn. Res. **11**, 2287–2322 (2010)
46. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (2013)
47. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
48. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends® Optim., 1(3):127–239 (2014)
49. Poliquin, R., Rockafellar, R.T.: Prox-regular functions in variational analysis. Trans. Am. Math. Soc. **348**(5), 1805–1838 (1996)
50. Poliquin, R., Rockafellar, R.T., Thibault, L.: Local differentiability of distance functions. Trans. Am. Math. Soc. **352**(11), 5231–5249 (2000)
51. Polyak, B.T.: Introduction to Optimization. Optimization Software, Cambridge (1987)
52. Rockafellar, R.T.: Characterizing firm nonexpansiveness of prox mappings both locally and globally. J. Nonlinear Convex Anal., **22**(5) (2021)

53. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis, vol. 317. Springer Science & Business Media, Berlin (2009)
54. Rudin, W.: Principles of Mathematical Analysis. McGraw-hill, New York (1986)
55. Ryu, E.K.: Uniqueness of DRS as the 2 operator resolvent-splitting and impossibility of 3 operator resolvent-splitting. Math. Program. **182**(1), 233–273 (2020)
56. Ryu, E.K., Boyd, S.: Primer on monotone operator methods. Appl. Comput. Math. **15**(1), 3–43 (2016)
57. Ryu, E.K., Yin, W.: Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators. Cambridge University Press, Cambridge (2022)
58. Saunderson, J., Chandrasekaran, V., Parrilo, P., Willsky, A.S.: Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. SIAM J. Matrix Anal. Appl. **33**(4), 1395–1416 (2012)
59. Shapiro, A.: Existence and differentiability of metric projections in Hilbert spaces. SIAM J. Optim. **4**(1), 130–141 (1994)
60. Srikumar, V., Manning, C.D.: Learning distributed representations for structured output prediction. Adv. Neural Inf. Process. Syst. **27** (2014)
61. Stella, L., Antonello, N., Fält, M., Volodin, D., Herceg, D., Saba, E., Carlson, F.B., Kelman, T., Brown, E., TagBot, J., Sopasakis, P.: JuliaFirstOrder/ProximalOperators.jl: v0.16.1. https://doi.org/10.5281/zenodo.10048760, (2023)
62. Takapoui, R.: *The Alternating Direction Method of Multipliers for Mixed-integer Optimization Applications*. PhD thesis, Stanford University (2017)
63. Takapoui, R., Moehle, N., Boyd, S., Bemporad, A.: A simple effective heuristic for embedded mixed-integer quadratic programming. Int. J. Control 1–11 (2017)
64. ten Berge, J.M.F.: Some recent developments in factor analysis and the search for proper communalities. In: Advances in Data Science and Classification, pp. 325–334. Springer, Berlin (1998)
65. Themelis, A., Patrinos, P.: Douglas-Rachford splitting and ADMM for nonconvex optimization: Tight convergence results. SIAM J. Optim. **30**(1), 149–181 (2020)
66. Tillmann, A.M, Bienstock, D., Lodi, A., Schwartz, A.: Cardinality minimization, constraints, and regularization: A survey. arXiv preprint arXiv:2106.09606 (2021)
67. Tropp, J.A.: Just relax: Convex programming methods for identifying sparse signals in noise. IEEE Trans. Inf. Theory (2006)
68. Udell, M., Horn, C., Zadeh, R., Boyd, S., et al.: Generalized low rank models. *Found. Trends® Mach. Learn.*, 9(1):1–118 (2016)
69. Vial, J.-P.: Strong and weak convexity of sets and functions. Math. Oper. Res. **8**(2), 231–259 (1983)