Computational Optimization and Applications (2024) 89:927–975
https://doi.org/10.1007/s10589-024-00607-2

Nonsmooth projection-free optimization with functional constraints

Kamiar Asgari[1] · **Michael J. Neely**[1]

Received: 18 November 2023 / Accepted: 2 September 2024 / Published online: 26 September 2024

Abstract

This paper presents a subgradient-based algorithm for constrained nonsmooth convex optimization that does not require projections onto the feasible set. While the well-established Frank–Wolfe algorithm and its variants already avoid projections, they are primarily designed for smooth objective functions. In contrast, our proposed algorithm can handle nonsmooth problems with general convex functional inequality constraints. It achieves an ϵ-suboptimal solution in $\mathcal{O}(\epsilon^{-2})$ iterations, with each iteration requiring only a single (potentially inexact) Linear Minimization Oracle call and a (possibly inexact) subgradient computation. This performance is consistent with existing lower bounds. Similar performance is observed when deterministic subgradients are replaced with stochastic subgradients. In the special case where there are no functional inequality constraints, our algorithm competes favorably with a recent nonsmooth projection-free method designed for constraint-free problems. Our approach utilizes a simple separation scheme in conjunction with a new Lagrange multiplier update rule.

Keywords Projection-free optimization · Frank–Wolfe method · Nonsmooth convex optimization · Stochastic optimization · Functional constraints

Mathematics Subject Classification 65K05 · 65K10 · 65K99 · 90C25 · 90C15 · 90C30

✉ Kamiar Asgari
Kamiaras@usc.edu

Michael J. Neely
Mjneely@usc.edu

[1] Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA

1 Introduction

Let $\mathbb{V}$ be a finite-dimensional real inner product space, such as $\mathbb{V} = \mathbb{R}^d$, for instance. Fix m as a nonnegative integer. This paper considers the problem

$$\begin{aligned} \text{Minimize:} \quad & f(x) \\ \text{Subject to:} \quad & h_i(x) \leq 0 \quad \forall i \in \{1, \ldots, m\} \\ & x \in \mathcal{X} \end{aligned}$$

where $f : \mathbb{V} \to \mathbb{R}$ and $h_i : \mathbb{V} \to \mathbb{R}$ for $i \in \{1, \ldots, m\}$ are convex continuous functions; $\mathcal{X} \subseteq \mathbb{V}$ is a compact and convex set. Such *convex optimization problems* have applications in fields such as machine learning, statistics, and signal processing [1–3]. While powerful numerical methods like the interior-point method and Newton's method are useful [4, 5], they can be computationally intensive for large problems with many dimensions (such as $\mathbb{V} = \mathbb{R}^d$ where d is large). This has prompted interest in *first-order methods* for large-scale problems [6, 7].

Many first-order methods solve subproblems that involve projections onto the feasible set $\mathcal{X}$. This projection step can be computationally expensive in high dimensions [8, 9]. To avoid this, some first-order methods replace the projection with a linear minimization over the set $\mathcal{X}$ [9–11]. For a given $v \in \mathbb{V}$ the Linear Minimization Oracle (LMO) over the set $\mathcal{X}$ returns a point $x \in \mathcal{X}$ such that:

$$x \in \arg\min\{\langle v\,,\, x\rangle : x \in \mathcal{X}\}.$$

The vast majority of such *projection-free* methods treat smooth objective functions and/or do not have functional inequality constraints [12–16]. Our paper considers a simple black-box method for general (potentially nonsmooth) convex objective and constraint functions. For a given $\epsilon > 0$, the method yields an approximate solution within $\mathcal{O}(\epsilon)$ of optimality with $\mathcal{O}(\epsilon^{-2})$ iterations, with each iteration requiring one (possibly inexact) subgradient calculation and one (possibly inexact) linear minimization over the set $\mathcal{X}$. This performance matches the existing lower bounds for the number of subgradient calculations in first-order methods, which may involve projections, and the number of linear minimizations for projection-free methods, as established in prior research [5, 17–20].

1.1 Prior work

The *Frank–Wolfe algorithm*, introduced in [13], pioneered the replacement of the projection step with a linear minimization. Initially, this approach was developed for problems with polytope domains. The Frank–Wolfe algorithm is also known as the *conditional gradient method* [15]. Variants of Frank–Wolfe have found application in diverse fields, including structured support vector machines [21], robust matrix recovery [22, 23], approximate Carathéodory problems [24], and reinforcement learning [25, 26]. Besides their notable computational efficiency achieved through avoiding computationally expensive projection steps, Frank–Wolfe-style algorithms offer an

additional advantage in terms of sparsity. This means that the algorithm iterates can be succinctly represented as convex combinations of several points located on the boundary of the relevant set. Such sparsity properties can be highly desirable in various practical applications [12, 27].

Most Frank–Wolfe-style algorithms are only designed for smooth objective functions. Some of these approaches handle functional inequality constraints by redefining the feasible set as the intersection of the set $\mathcal{X}$ and the functional constraints, potentially eliminating the computational advantages of linear minimization over the feasible set by changing it. Extending these methods to cope with nonsmooth objective and constraint functions is far from straightforward. A simple two-dimensional example in [28] shows how convergence can fail when the basic Frank–Wolfe algorithm is used for nonsmooth problems (replacing gradients with subgradients).

Initial efforts to extend Frank–Wolfe to nonsmooth problems can be found in [29–31]. These methods require analytical preparations for the objective function and are applicable to specific function classes. They are distinct from black-box algorithms that work for general problems.

Another idea, initially introduced by [14] and later revisited by [16], involves smoothing the nonsmooth objective function using a Moreau envelope [32]. This approach demands access to a *proximity operator* associated with the objective function. While some nonsmooth functions have easily solvable proximity operators [33], many do not. In general, the worst-case complexity of a single proximal iteration can be the same as the complexity of solving the original optimization problem [34]. An alternative concept presented in [35] uses $\mathcal{O}(\epsilon^{-2})$ queries to a Fenchel-type oracle. However, the Fenchel-type oracle is only straightforward to implement for specific classes of nonsmooth functions.

Another approach, proposed by [17], utilizes random smoothing (for a general analysis of random smoothing, see [36]). This method requires $\mathcal{O}(\epsilon^{-2})$ queries to an LMO, which was proven to be optimal in the same work [17]. Unlike the previously mentioned methods, this algorithm only relies on access to a first-order oracle. However, it falls short in terms of the number of calls to the first-order oracle $\mathcal{O}(\epsilon^{-4})$ compared to the optimal $\mathcal{O}(\epsilon^{-2})$ achieved by projected subgradient descent [18, 19]).

In an effort to adapt the Frank–Wolfe algorithm to an online setting, [37] successfully achieved a convergence rate of $\mathcal{O}(\epsilon^{-3})$ for both offline and stochastic optimization problems with nonsmooth objective function. This was accomplished with just one call to an LMO in each round.

In the context of projection-free methods for nonsmooth problems, the work [38] was the first to achieve optimal $\mathcal{O}(\epsilon^{-2})$ query complexity for both the LMO and the first-order oracle that obtains subgradients. This was made possible through the idea of approximating the Moreau envelope.

Our current paper introduces a different approach to achieve $\mathcal{O}(\epsilon^{-2})$ query complexity. In the special case of problems without functional inequality constraints, it competes favorably with the work [38]. Moreover, our algorithm distinguishes itself by its ability to handle functional inequality constraints, a feature not present in [38].

There are many algorithms (though not necessarily in a projection-free manner) that address convex optimization with general nonsmooth objectives and constraint functions. These prior works explore various techniques, including cooperative sub-

gradients [39, 40], the level-set method [5, 41, 42], reformulation as saddle point problems [43], exact penalty and augmented Lagrangian methods [44–47], Lyapunov drift-plus-penalty [48–50], bundle and fiber methods [51, 52], and constraint extrapolation [53].

There have been several efforts to generalize projection-free algorithms to handle more than one set constraint. The Frank–Wolfe algorithm has been extended to stochastic affine constraints in [54]. More recently, [55, 56] have developed projection-free methods for problems with functional constraints. However, unlike our approach, which assumes no special properties of the functions, all mentioned methods assume the objective and constraint functions are either smooth or structured nonsmooth.

1.1.1 Other projection-free methods

The predominant body of literature on projection-free methods, including this paper, typically assumes the existence of a Linear Minimization Oracle (LMO) for the feasible set $\mathcal{X}$. However, recent alternative approaches in [57–67] utilize various techniques, such as separation Oracles, membership Oracles, Newton iterations, and radial dual transformations. It is worth noting that some of these oracles can be implemented using others, as demonstrated, for instance, in [59]. Nevertheless, none of these approaches can be considered universally superior to others in terms of implementation efficiency.

1.2 Our contribution

This paper introduces a projection-free algorithm designed for general convex optimization problems, with both feasible set and functional constraints. Our approach has mathematical guarantees to work where both the objective and constraint functions are nonsmooth, relying on access to only possibly inexact subgradient oracles for these functions. While previous projection-free methods in the literature have engaged with similar optimization challenges, they have primarily not included functional constraints or have been limited to smoothable nonsmooth functions. To the best of our knowledge, our algorithm is the first to address this category of problems in a projection-free manner comprehensively.

Our algorithm achieves an optimal performance of $\mathcal{O}(\epsilon^{-2})$, notably even in scenarios where the LMO exhibits imprecision. This aspect is particularly crucial considering that for certain sets, the inexact LMO offers the computational advantage over projection onto those sets (for example, see [12, 68]).

The derivation of our algorithm is notably distinct, as it more closely resembles subgradient-descent-type algorithms rather than those of the Frank–Wolfe-type. We start with a simple separation idea that enables each iteration to be separated into: (i) A linear minimization over the feasible set $\mathcal{X}$; (ii) A projection onto a much simpler set $\mathcal{Y} \subseteq \mathbb{V}$ (this includes using $\mathcal{Y} = \mathbb{V}$, for which the projection step is trivial).[1] This separation sets the stage for a unique *Lagrange multiplier update rule* of the form

[1] See Appendix D for more discussion on choosing the set $\mathcal{Y}$.

$$W_{i,t+1} = \max\left\{W_{i,t} + h_i(y_t) + \left\langle h_i'(y_t)\,,\ y_{t+1} - y_t\right\rangle, \left[-h_i(y_{t+1})\right]_+\right\}.$$

Traditional Lagrange multiplier updates replace the right-hand-side with a maximum with 0, rather than a maximum with $\left[-h_i(y_{t+1})\right]_+$ (see, for example, the classic update rule for the dual subgradient algorithm in [1, 69, 70]). Our update is inspired by a related update used in [46] for a different class of problems. However, the update in [46] takes a max with $-h_i(y_{t+1})$ rather than its positive part. Our approach has advantages in the projection-free scenario and may have applications in other settings.

1.3 Applications

The proposed algorithm may be useful in problems where constraints are linear, and the objective function is nonsmooth. For example, in network optimization, the constraints model channel capacity limits, which can be expressed as linear inequalities and equalities [71, 72]. The directed graph structure of the network establishes a flow polytope constraint; consequently, minimizing a linear objective over this set involves identifying the minimum weight path based on the assigned edge weights. The flow polytope is one of the sets where linear minimization is significantly cheaper than projection [68, 73]. The objective function, which describes utility and fairness [74, 75], can be nonsmooth due to piecewise linearities and/or being the maximum of multiple convex functions.

Our algorithm also holds potential for application in Quantum State Tomography (QST), which presents a nonsmooth and stochastic problem. Earlier work employing Frank–Wolfe-type methods for this issue has utilized smoothing techniques [76].

Robust Structural Risk Minimization is another class of problems that can benefit from our algorithm, mainly when sparsity is crucial. Our algorithm is well-equipped to handle the inherent stochastic characteristics of the problem and the nonsmooth nature of loss functions like the l_1-norm or the Hinge function, which are commonly utilized for robustness purposes [77–80]. Furthermore, the algorithm is adaptable to complex scenarios such as Fused Lasso regression [81], enforcing additional desired structures through functional constraints.

1.4 Notation

The set of nonnegative real numbers are denoted as $\mathbb{R}_+ \subseteq \mathbb{R}$. Our underlying space for optimization is denoted as $\mathbb{V}$ and is assumed to be a finite-dimensional inner product space with a general inner product $\langle v\,,\ u\rangle$ and a norm determined by the inner product, i.e., $\|\cdot\| = \sqrt{\langle\cdot\,,\ \cdot\rangle}$. Our examples consider $\mathbb{V} = \mathbb{R}^d$ with inner product given by the dot product $\langle v\,,\ u\rangle := v^\top u$, and $\mathbb{V} = \mathbb{R}^{q\times p}$ (for matrices) with inner product $\langle v\,,\ u\rangle := \mathrm{Tr}\left(v^\top u\right)$. The positive part of a real number x is denoted $[x]_+ := \max\{0, x\}$ and is also applied element-wise for elements of $\mathbb{R}^m$. The subdifferential of a function f at point x is denoted by $\partial f(x)$, with $f'(x)$ representing a particular (arbitrary) subgradient of f at x.

2 Formulation and problem separation

For a finite-dimensional inner product space $\mathbb{V}$ and a compact set $\mathcal{X} \subseteq \mathbb{V}$:

$$\begin{aligned} \text{Minimize:} \quad & f(x) \\ \text{Subject to:} \quad & h_i(x) \leq 0 \quad \forall i \in \{1, \ldots, m\} \\ & x \in \mathcal{X} \end{aligned} \tag{P1}$$

where $f : \mathbb{V} \to \mathbb{R}$ and $h_i : \mathbb{V} \to \mathbb{R}$ for $i \in \{1, \ldots, m\}$ are proper, continuous, convex functions. Let $\mathcal{X}^* \subseteq \mathcal{X}$ be the set of optimal solutions. It is assumed that $\mathcal{X}^*$ is nonempty. Let f^* represent the optimal objective value. It follows by compactness of $\mathcal{X}$ that f^* is finite.

The primary goal is to find an ϵ-suboptimal solution to Problem (P1). This can involve numerical steps that make use of oracles that return random vectors. The output of the algorithm is the construction of a random vector $\bar{x} \in \mathcal{X}$ such that

$$\mathbb{E}\{f(\bar{x})\} - f^* \leq \mathcal{O}(\epsilon),$$

and such that

$$\mathbb{E}\left\{\| [h(\bar{x})]_+ \|_2\right\} \equiv \mathbb{E}\left\{\sqrt{\sum_{i=1}^{m} (\max\{0,\ h_i(\bar{x})\})^2}\right\} \leq \mathcal{O}(\epsilon),$$

where $h(x) = (h_1(x), \ldots, h_m(x))^\top$ and $\|\cdot\|_2$ refers to the standard l_2-norm defined on vector space $\mathbb{R}^m$. When the oracles are deterministic the expectations can be removed.

Assumption 1 The feasible set $\mathcal{X}$ is a compact convex subset of $\mathbb{V}$, and there is a known bound D on the diameter of the set $\mathcal{X}$, such that

$$\|x - y\| \leq D \quad \forall x, y \in \mathcal{X}$$

Assumption 2 There exists a vector (Lagrange multiplier) $\mu \in \mathbb{R}^m_+$ such that:

$$f^* \leq f(x) + \mu^\top h(x) \quad \forall x \in \mathcal{X}. \tag{1}$$

Assumption 3 The algorithm has access to the following computation oracles:

3.i Inexact Linear Minimization Oracle (In- LMO):[2] Given a unit vector $v \in \mathbb{V}$ and a desired error upper bound $\delta \geq 0$, this oracle returns a random point $x \leftarrow \text{In- LMO}_{\mathcal{X}}\{v; \delta\}$ such that x belongs to the set $\mathcal{X}$ and

$$\mathbb{E}\{\langle x\,,\ v\rangle\} \leq \langle y\,,\ v\rangle + \delta \quad \forall y \in \mathcal{X}$$

[2] This oracle is formulated similarly to the approximate oracle described in [12]. Note that for a fixed δ, the computational cost of running $\text{In- LMO}_{\mathcal{X}}\{v; \delta\}$ may increase with the size of v. For further discussion, see Appendix C.

3.ii Projection Oracle (PO): There is a closed convex set $\mathcal{Y} \subseteq \mathbb{V}$ such that $\mathcal{X} \subseteq \mathcal{Y}$. Given $v \in \mathbb{V}$, this oracle returns a point $\text{PO}_{\mathcal{Y}}\{v\} \in \mathcal{Y}$ such that:

$$\text{PO}_{\mathcal{Y}}\{v\} := \arg\min_{y \in \mathcal{Y}} \|v - y\|. \tag{2}$$

3.iii Stochastic Subgradient Oracle: Given $y \in \mathcal{Y}$, this oracle independently returns $m + 1$ random vectors $s, g_1, \ldots, g_m$ such that

$$\mathbb{E}\left\{s \middle| y\right\} \in \partial f(y),$$
$$\mathbb{E}\left\{g_i \middle| y\right\} \in \partial h_i(y) \quad \forall i \in \{1, \ldots, m\}.$$

Assume there are known real-valued constants $L, G \geq 0$ and unknown real-valued constants $G_1, \ldots, G_m \geq 0$ such that:[3]

$$\|g_i\| \leq G_i \quad \forall i \in \{1, \ldots, m\}$$
$$\sum_{i=1}^{m} G_i^2 \leq G^2$$
$$\sqrt{\mathbb{E}\left\{\|s\|^2 \middle| y\right\}} \leq L$$

so that the g_i vectors are deterministically bounded while the s vector is required only to have a finite second moment. It is worth noting that by the law of iterated expectation, we obtain: $\sqrt{\mathbb{E}\left\{\|s\|^2\right\}} \leq L$.

3.iv Function Value Oracle: This oracle takes a point $y \in \mathcal{Y}$ and provides the values $h_i(y)$ for $i \in \{1, \ldots, m\}$ as its output.

Assumption 4 The algorithm has access to an initial point belonging to the set $\mathcal{X}$, specifically, $x_1 \in \mathcal{X}$.

Remark 1 To satisfy Assumption 4, a straightforward selection for x_1 is $x_1 \leftarrow$ IN- $\text{LMO}_{\mathcal{X}}\{v; \delta\}$, where v could be as simple as $v = \mathbf{0}$. Nevertheless, x_1 could also be any other point within $\mathcal{X}$ that the practitioner considers to be closer to the optimal set $\mathcal{X}^*$.

Definition 1 Let $\mathbb{V}$ and $\mathbb{V}'$ be two vector spaces endowed with their respective norms $\|\cdot\|$ and $\|\cdot\|'$. A function $r : \mathcal{Y} \to \mathbb{V}'$ is termed Lipschitz continuous over the set $\mathcal{Y} \subseteq \mathbb{V}$ with a Lipschitz constant $\zeta > 0$ if it satisfies the condition that for every pair of points x and y in $\mathcal{Y}$, the following inequality holds:

$$\|r(x) - r(y)\|' \leq \zeta \|x - y\|.$$

[3] This assumption may fail in some cases. See Appendix D for further remarks.

Lemma 1 *If Assumption 3.iii is met, then the functions $f : \mathbb{V} \to \mathbb{R}$, $h : \mathbb{V} \to \mathbb{R}^m$, and $h_i : \mathbb{V} \to \mathbb{R}$ (for all $i \in \{1, \ldots, m\}$) demonstrate Lipschitz continuity over the set $\mathcal{Y}$ with Lipschitz constants not exceeding L, G, and G_i, respectively.*

Proof See Appendix A. □

2.1 Problem separation

Recall that $\mathcal{X} \subseteq \mathcal{Y}$. It is clear that Problem (P1) is equivalent to

$$\begin{aligned} \text{Minimize:} \quad & f(y) \\ \text{Subject to:} \quad & h_i(y) \leq 0 \quad \forall i \in \{1, \ldots, m\} \\ & y = x \\ & x \in \mathcal{X} \quad ; y \in \mathcal{Y}. \end{aligned} \tag{P2}$$

Problem (P2) is said to have a Lagrange multiplier vector (μ, λ), where $\mu \in \mathbb{R}^m_+$ and $\lambda \in \mathbb{V}$, if

$$f^* \leq f(y) + \mu^\top h(y) + \langle \lambda, x - y \rangle \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{3}$$

Note that the right-hand-side of the above inequality uses the general inner product in $\mathbb{V}$ for describing the contribution of the λ multiplier. The next lemma shows that the new Problem (P2) has Lagrange multipliers whenever the original Problem (P1) does, and the new multipliers can be described in terms of the originals. The key connection between the two problems arises by considering subgradients of the convex function $v : \mathbb{V} \to \mathbb{R}$ defined by

$$v(x) = f(x) + \mu^\top h(x) \quad \forall x \in \mathbb{V}. \tag{4}$$

where μ is a Lagrange multiplier of Problem (P1). Note that the real-valued convex functions f, h_i, v have domains equal to the entire space $\mathbb{V}$.

Lemma 2 *(Lagrange Multipliers) Suppose the original Problem (P1) has a Lagrange multiplier $\mu \in \mathbb{R}^m_+$ (so that Assumption 2 holds), and further assume Assumption 3.iii is satisfied. Fix $x^* \in \mathcal{X}^*$. Then there exists a $\lambda \in \mathbb{V}$ such that the pair (μ, λ) forms a Lagrange multiplier for Problem (P2), meaning that (3) holds, and additionally satisfies:*

$$\|\lambda\| \leq L + G\|\mu\|_2. \tag{5}$$

Proof Since μ is a Lagrange multiplier of the original Problem (P1), we have, by (1) and the definition of function v:

$$v(x) \geq f^* \quad \forall x \in \mathcal{X}. \tag{6}$$

Applying (6) to the point $x^* \in \mathcal{X}$ gives

$$\begin{aligned} f^* &\leq v(x^*) \\ &\overset{(a)}{=} f(x^*) + \mu^\top h(x^*) \\ &= f^* + \mu^\top h(x^*) \\ &\overset{(b)}{\leq} f^* \end{aligned}$$

where (a) holds by definition of v in (4); (b) holds because $\mu \geq 0$ and $h(x^*) \leq 0$ (where these vector inequalities are taken entrywise). The above chain of inequalities simultaneously proves:

$$v(x^*) = f^* \tag{7}$$

$$\mu^\top h(x^*) = 0 \tag{8}$$

Equality (7) together with (6) implies that x^* minimizes the convex function $v : \mathbb{V} \to \mathbb{R}$ over the restricted set of all $x \in \mathcal{X}$. Thus, Prop B.24f from [44] ensures *there exists* a subgradient $\lambda \in \partial v(x^*)$ that satisfies:

$$\langle \lambda \,,\, x - x^* \rangle \geq 0 \quad \forall x \in \mathcal{X}. \tag{9}$$

(The property (9) is not necessarily satisfied by *all* subgradients in $\partial v(x^*)$). Fix $y \in \mathbb{V}$ and $x \in \mathcal{X}$. Since $\lambda \in \partial v(x^*)$ we have, by the definition of a subgradient:

$$v(y) \geq v(x^*) + \langle \lambda \,,\, y - x^* \rangle$$

Substituting the definition of v in (4) into the above inequality gives

$$\begin{aligned} f(y) + \mu^\top h(y) &\geq f(x^*) + \mu^\top h(x^*) + \langle \lambda \,,\, y - x^* \rangle \\ &\overset{(a)}{=} f^* + \langle \lambda \,,\, y - x^* \rangle \\ &= f^* + \langle \lambda \,,\, y - x \rangle + \langle \lambda \,,\, x - x^* \rangle \\ &\overset{(b)}{\geq} f^* + \langle \lambda \,,\, y - x \rangle \end{aligned}$$

where (a) holds by (8); (b) holds by (9). This holds for all $y \in \mathbb{V}$ and $x \in \mathcal{X}$. Since $\mathcal{Y} \subseteq \mathbb{V}$, it certainly holds for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. This proves the desired Lagrange multiplier inequality (3).

This particular $\lambda \in \partial v(x^*)$ has the form

$$\lambda = f'(x^*) + \sum_{i=1}^{m} \mu_i h_i'(x^*), \tag{10}$$

for some particular subgradients in $\partial f(x^*)$ and $\partial h_i(x^*)$ for $i \in \{1, \ldots, m\}$. This follows by the fact that v is a sum of convex functions and hence $\partial v(x^*)$ is the

Minkowski sum of the subdifferentials of those component functions (see, for example, Prop B.24b [44]).

Taking the norm of both sides of (10) and applying the triangle inequality (noting that $\mu_i \geq 0$), we obtain:

$$\|\lambda\| = \left\| f'(x^*) + \sum_{i=1}^{m} \mu_i h_i'(x^*) \right\| \leq \left\| f'(x^*) \right\| + \sum_{i=1}^{m} \mu_i \left\| h_i'(x^*) \right\|$$

Adding the Cauchy-Schwarz inequality, we get

$$\|\lambda\| \leq \left\| f'(x^*) \right\| + \|\mu_i\|_2 \sqrt{\sum_{i=1}^{m} \left\| h_i'(x^*) \right\|^2} \tag{11}$$

Here we need to consider two cases:

i If x^* belongs to the interior of the set $\mathcal{Y}$, then Lipschitz continuity of f and h_i proved in Lemma 1 implies that (see, for example, part (ii) of Theorem 3.61 [6]):

$$\sum_{i=1}^{m} \left\| h_i'(x^*) \right\|^2 \leq G^2$$
$$\left\| f'(x^*) \right\| \leq L$$

which concludes the proof.

ii If x^* does not belong to the interior of the set $\mathcal{Y}$, then we cannot directly use Lipschitz continuity to get a bound of the subgradients. The reason is that the Lipschitz continuity of a function over $\mathcal{Y}$ does not guarantee the boundedness of every subgradient by the Lipschitz constant. We employ the *McShane-Whitney extension theorem* [82] to overcome this. Part of this theorem establishes that if $r : \mathcal{Y} \to \mathbb{R}$ is a convex and ζ-Lipschitz continuous function defined on the convex set $\mathcal{Y}$, then there exists an extended convex function $\tilde{r} : \mathbb{V} \to \mathbb{R}$ which satisfies the following conditions:

(a) $r(x) = \tilde{r}(x)$ for all $x \in \mathcal{Y}$.
(b) For any $x \in \mathbb{V}$, all subgradients $s \in \partial\tilde{r}(x)$ have $\|s\| \leq \zeta$.

By part (a) of this theorem, our proof until (11) can be stated using the extended functions $\tilde{f}$ and $\tilde{h}$. Thus, we can conclude that there exists a $\lambda \in \partial(\tilde{f} + \mu^\top \tilde{h})(x^*)$ such that the pair (μ, λ) forms a Lagrange multiplier for Problem (P2), and this particular λ has the form

$$\lambda = \tilde{f}'(x^*) + \sum_{i=1}^{m} \mu_i \tilde{h}_i'(x^*),$$

for some particular subgradients in $\partial\tilde{f}(x^*)$ and $\partial\tilde{h}_i(x^*)$ for $i \in \{1, \ldots, m\}$. Part (b) of the theorem implies that the functions $\tilde{f} : \mathbb{V} \to \mathbb{R}$ and $\tilde{h}_i : \mathbb{V} \to \mathbb{R}$ (for

all $i \in \{1, \ldots, m\}$) demonstrate Lipschitz continuity over the set $\mathbb{V}$, including the boundary of the set $\mathcal{X}$, with Lipschitz constants not exceeding L and G_i, respectively. Thus,

$$\sum_{i=1}^{m} \left\| \tilde{h}_i'(x^*) \right\|^2 \leq G^2,$$
$$\left\| \tilde{f}'(x^*) \right\| \leq L.$$

This concludes the proof. □

2.2 Algorithm intuition

The new Problem (P2) uses two decision variables $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. This is useful precisely because of the Lagrange multiplier result (3). Our approach is as follows: First imagine that we know the Lagrange multipliers μ and λ. Suppose we seek to minimize the right-hand-side of (3) over all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. This separates into two subproblems:

- Chose $x \in \mathcal{X}$ to minimize the *linear function* $\langle \lambda\,,\, x \rangle$. This is done (in a possibly inexact way) by the oracle IN- $\text{LMO}_{\mathcal{X}}$.
- Choose $y \in \mathcal{Y}$ to minimize the *possibly nonsmooth convex function* $f(y) + \mu^\top h(y) - \langle \lambda\,,\, y \rangle$. This is done by using subgradients and projecting onto the set $\mathcal{Y}$ via the oracle $\text{PO}_{\mathcal{Y}}$. The set $\mathcal{Y}$ is chosen to be a set that contains $\mathcal{X}$. Further, $\mathcal{Y}$ is assumed to have a structure that is very simple so that projections onto $\mathcal{Y}$ are easy. For instance, if $\mathcal{Y}$ is a box, a norm-ball of fixed radius centered at the origin, or the entire space $\mathbb{V}$ itself, then the projections are straightforward. Since we avoid complicated projections onto the feasible set $\mathcal{X}$, our algorithm is "projection-free".

Of course, the Lagrange multipliers μ and λ are unknown. Therefore, our algorithm must use approximations of these multipliers that are updated as time goes on. Further, even if μ and λ were known, minimizing the right-hand-side of (3) may not have a desirable result. That is because the right-hand-side of (3) may have many minimizers, not all of them satisfying the desired constraints. Therefore, our update rule is carefully designed to ensure convergence to a vector that satisfies the desired constraints.

3 The new algorithm

We call our algorithm Nonsmooth Projection-Free Optimization with Functional Constraints. This algorithm uses a parameter $T \in \{1, 2, 3, \ldots\}$ (which determines the number of iterations) and additional parameters $\eta > 0, \alpha > 0, \beta > 0$. We focus on two specific parameter choices:

- Parameter Selection 1: Fix $\epsilon > 0$ and define

$$\eta = \epsilon, \quad \alpha = \beta = 1/\epsilon, \quad T \geq 1/\epsilon^2 \qquad \text{(ParSel.1)}$$

Algorithm 1 Nonsmooth Projection-Free Optimization with Functional Constraints (Nonsmooth PF-FC)

Require: Parameters: $T, \eta, \alpha, \beta, \delta$. Initial point: $x_1 \in \mathcal{X}$.

$y_1 \leftarrow x_1$

$Q_1 \leftarrow \mathbf{0}$

Obtain stoch subgradients at y_1: $s_1, g_{1,1}, \ldots, g_{m,1}$

$W_1 \leftarrow [-h(y_1)]_+$

for $1 \leq t \leq T-1$ **do**

  $x_{t+1} \leftarrow \text{IN- LMO}_{\mathcal{X}}\{-Q_t; \delta\}$

  $p_t \leftarrow \eta Q_t + s_t + \beta \sum_{i=1}^{m} (W_{i,t} + h_i(y_t)) g_{i,t}$

  $\tilde{y}_{t+1} \leftarrow \dfrac{\left(\alpha + 2G^2\beta\right) y_t + \eta x_{t+1} - p_t}{\alpha + 2G^2\beta + \eta}$

  $y_{t+1} \leftarrow \text{PO}_{\mathcal{Y}}\{\tilde{y}_{t+1}\}$

  ▷ `Lines 7,8 and 9 are only separated for better readability.`

  $Q_{t+1} \leftarrow Q_t + y_{t+1} - x_{t+1}$

  Obtain stoch subgradients at y_{t+1}: $s_{t+1}, g_{1,t+1}, \ldots, g_{m,t+1}$

  for $1 \leq i \leq m$ **do**

    $W_{i,t+1} \leftarrow \max\left\{W_{i,t} + h_i(y_t) + \langle g_{i,t}\,,\, y_{t+1} - y_t\rangle, \left[-h_i(y_{t+1})\right]_+\right\}$

  end for

end for

return $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$

- Parameter Selection 2: Fix $T \in \{1, 2, 3, ...\}$ and define

$$\alpha = \frac{L\sqrt{T}}{D}, \quad \eta = \frac{L}{\sqrt{T\left(D^2 + 2\delta\right)}}, \quad \beta = \frac{\sqrt{T}}{GD} \tag{ParSel.2}$$

The first parameter selection is useful when the values D, L, δ associated with the problem structure are unknown. The second is fine tuned with knowledge of these values.

Remark 2 While the parameter δ appears to be just one of the parameters within the algorithm's configuration, it fundamentally differs from the others. The following theorem provides the trade-off between the error in the algorithm's output and the number of iterations T. It further demonstrates that if δ is maintained at an order of D^2, the main algorithm's error does not substantially increase.

This implies that there exists an optimizable trade-off in choosing δ: by increasing δ, we will need a larger T to achieve the same final error ϵ, but each iteration will be completed more quickly.

This is particularly important in practice, where the actual computational cost, rather than just the number of iterations T, is what matters. See Appendix C for more details.

Theorem 1 *(Objective gap) Given Assumptions 1–4, for Algorithm 1 with any* $T \in \{1, 2, 3, ...\}$, $\eta > 0$, $\alpha > 0$, $\beta > 0$, *and* $\delta \geq 0$ *the expected gap in the objective*

function is bounded as follows:

$$\mathbb{E}\{f(\bar{x}_T)\} - f(x^*) \leq \frac{L^2}{2T\eta} + \eta\frac{D^2+2\delta}{2} + \frac{L^2}{2\alpha} + \frac{\alpha D^2}{2T} + \frac{G^2D^2\beta}{T}$$

In particular, under Parameter Selection (ParSel.1) we have

$$\mathbb{E}\{f(\bar{x}_T)\} - f^* \leq \mathcal{O}(\epsilon) \quad \forall T \geq 1/\epsilon^2,$$

while under Parameter Selection (ParSel.2) we have

$$\mathbb{E}\{f(\bar{x}_T)\} - f^* \leq \left(L\sqrt{D^2+2\delta} + LD + GD\right)\frac{1}{\sqrt{T}}.$$

Theorem 2 *(Constraint violation) Given Assumptions 1–4, Algorithm 1 under Parameter Selection (ParSel.1) yields*

$$\mathbb{E}\left\{\| [h(\bar{x}_T)]_+ \|_2\right\} \leq \mathcal{O}(\epsilon) \quad \forall T \geq 1/\epsilon^2,$$

while under Parameter Selection (ParSel.2) we have

$$\mathbb{E}\left\{\| [h(\bar{x}_T)]_+ \|_2\right\} \leq \frac{1}{\sqrt{T}}\sqrt{A_0 + A_1\|\mu\|_2 + A_2\|\mu\|_2^2}.$$

Here, A_1, A_2, and A_3 are constants depending on the problem's constants (they are defined in the last part of the theorem's proof). The variable μ represents the Lagrange multiplier from (1).

The proof of the first theorem is provided in this section. The proof of the second theorem is in Appendix A.1.

Remark 3 When the number of iterations T is on the order of $\mathcal{O}(\epsilon^{-2})$, the expected suboptimality $\mathbb{E}\{f(\bar{x}_T)\} - f^*$ is bounded by $\mathcal{O}(\epsilon)$. This approach achieves an optimal solution in terms of the computational cost, measured by the number of calls to both the In- LMO and the (possibly stochastic) first-order oracle [5, 17, 19].

3.1 Lagrange multiplier update analysis

Line 10 of Algorithm 1 specifies that $Q_{t+1} = Q_t + y_{t+1} - x_{t+1}$. If we apply the $\|\cdot\|^2$ norm to both sides of this equation for all $t \in \{1, \ldots, T-1\}$, we obtain:

$$\langle Q_t\,,\, y_{t+1} - x_{t+1}\rangle + \frac{1}{2}\|y_{t+1} - x_{t+1}\|^2 = \frac{1}{2}\|Q_{t+1}\|^2 - \frac{1}{2}\|Q_t\|^2. \tag{12}$$

Furthermore, summing $Q_{t+1} = Q_t + y_{t+1} - x_{t+1}$ over t in the range $t \in \{1, \ldots, T\}$ and using Line 2 which states $Q_1 = 0$, gives:

$$Q_T = \sum_{t=1}^{T} (y_t - x_t) = T(\bar{y}_T - \bar{x}_T). \tag{13}$$

Here, similar to $\bar{x}_T$, we define $\bar{y}_T := \frac{1}{T} \sum_{t=1}^{T} y_t$.

For simplicity, define $l_{i,t}(x)$ as the linearization of h_i at the point y_t obtained from the algorithm. This linearization uses the stochastic subgradient $g_{i,t}$:

$$l_{i,t}(x) := h_i(y_t) + \langle g_{i,t} \, , \, x - y_t \rangle. \tag{14}$$

Define $l_t(x)$ as a vector, where each element at index i corresponds to $l_{i,t}(x)$.

Lemma 3 *Under Algorithm 1 with any $T \in \{1, 2, 3, \ldots\}$, $\eta > 0, \alpha > 0, \beta > 0$, we have for any $x^* \in \mathcal{X}^*$, $i \in \{1, \ldots, m\}$, and $t \in \{1, \ldots, T\}$*

$$W_{i,t} \geq 0 \tag{15}$$

$$W_{i,t} + h_i(y_t) \geq 0 \tag{16}$$

$$\mathbb{E}\left\{ (W_t + h(y_t))^\top l_t(x^*) \right\} \leq 0 \tag{17}$$

Further, for all $t \in \{1, \ldots, T-1\}$ we have

$$\begin{aligned} &(W_t + h(y_t))^\top l_t(y_{t+1}) + \frac{G^2}{2} \|y_{t+1} - y_t\|^2 \\ \geq &\frac{\|W_{t+1}\|_2^2}{2} - \frac{\|W_t\|_2^2}{2} - \frac{\|\left[-h(y_{t+1})\right]_+\|_2^2}{2} + \frac{\|h(y_t)\|_2^2}{2}. \end{aligned} \tag{18}$$

Proof Lines 4 at $t = 1$ and 13 at $t \geq 2$ establish that $W_{i,t} \geq \max\{0, -h_i(y_t)\}$, thereby confirming the validity of Eqs. (15) and (16).

Using the definition of the function l_t in Eq. (14) we have:

$$(W_t + h(y_t))^\top l_t(x^*) = \sum_{i=1}^{m} \left(W_{i,t} + h_i(y_t)\right) \left(h_i(y_t) + \langle g_{i,t} \, , \, x^* - y_t \rangle\right).$$

Using the iterated expectation gives:

$$\begin{aligned} &\mathbb{E}\left\{ \sum_{i=1}^{m} \left(W_{i,t} + h_i(y_t)\right) \left(h_i(y_t) + \langle g_{i,t} \, , \, x^* - y_t \rangle\right) \right\} \\ =&\mathbb{E}\left\{ \sum_{i=1}^{m} \mathbb{E}\left\{ W_{i,t} + h_i(y_t) \middle| y_t \right\} \left(h_i(y_t) + \langle \mathbb{E}\left\{ g_{i,t} \middle| y_t \right\} \, , \, x^* - y_t \rangle\right) \right\}. \end{aligned}$$

Here, we used the independence of $W_{i,t}$ and $g_{i,t}$ when conditioning on y_t. Assumption 3.iii implies $\mathbb{E}\left\{g_{i,t}\middle|y_t\right\} \in \partial h_i(y_t)$. Using Eq. (16) and convexity of the function h_i we get:

$$\begin{aligned}&\mathbb{E}\left\{W_{i,t}+h_i(y_t)\middle|y_t\right\}\left(h_i(y_t)+\left\langle\mathbb{E}\left\{g_{i,t}\middle|y_t\right\}\,,\,x^*-y_t\right\rangle\right)\\ \leq&\mathbb{E}\left\{W_{i,t}+h_i(y_t)\middle|y_t\right\}h_i(x^*)\end{aligned}$$

Using the inequality $h_i(x^*) \leq 0$ proves Eq. (17).

Applying the inequality $(\max\{a,b\})^2 \leq a^2+b^2$ to Line 13 gives

$$\begin{aligned}\frac{W_{i,t+1}^2}{2} \leq&\frac{W_{i,t}^2}{2}+W_{i,t}\left(h_i(y_t)+\left\langle g_{i,t}\,,\,y_{t+1}-y_t\right\rangle\right)\\ &+\frac{1}{2}\left(h_i(y_t)+\left\langle g_{i,t}\,,\,y_{t+1}-y_t\right\rangle\right)^2+\frac{\left[-h_i(y_{t+1})\right]_+^2}{2}\\ =&\frac{W_{i,t}^2}{2}+W_{i,t}\left(h_i(y_t)+\left\langle g_{i,t}\,,\,y_{t+1}-y_t\right\rangle\right)\\ &+h_i(y_t)\left(h_i(y_t)+\left\langle g_{i,t}\,,\,y_{t+1}-y_t\right\rangle\right)\\ &-\frac{(h_i(y_t))^2}{2}+\frac{1}{2}\left(\left\langle g_{i,t}\,,\,y_{t+1}-y_t\right\rangle\right)^2+\frac{\left[-h_i(y_{t+1})\right]_+^2}{2}\\ \leq&\frac{W_{i,t}^2}{2}+\left(W_{i,t}+h_i(y_t)\right)\left(h_i(y_t)+\left\langle g_{i,t}\,,\,y_{t+1}-y_t\right\rangle\right)\\ &+\frac{\|g_{i,t}\|^2}{2}\|y_{t+1}-y_t\|^2+\frac{\left[-h_i(y_{t+1})\right]_+^2}{2}-\frac{(h_i(y_t))^2}{2}.\end{aligned}$$

Here, the final step utilizes Cauchy-Schwarz inequality. By summing over the range $i \in \{1,2,\ldots,m\}$ and leveraging Assumption 3.iii, which implies $\sum_{i=1}^m \|g_{i,t}\|^2 \leq G^2$, and using the linearized notation (14), we get the proof of Eq. (18). □

3.2 Algorithm analysis

By definition of s_t as a stochastic subgradient of f at y_t we have

$$\mathbb{E}\left\{\left\langle s_t\,,\,x^*-y_t\right\rangle\middle|y_t\right\} \leq f(x^*)-f(y_t).$$

By iterated expectations we have

$$\mathbb{E}\left\{\left\langle s_t\,,\,x^*-y_t\right\rangle\right\} \leq f(x^*)-\mathbb{E}\left\{f(y_t)\right\}. \tag{19}$$

Line 6 of Algorithm 1 gives $x_{t+1} \leftarrow \text{IN-LMO}_{\mathcal{X}}\{-Q_t;\delta\}$. Since we have $x^* \in \mathcal{X}$, Assumption 3.i ensures that x_{t+1} satisfies

$$\mathbb{E}\left\{\left\langle x_{t+1}\,,\,-Q_t\right\rangle\middle|Q_t\right\} \leq \left\langle x^*\,,\,-Q_t\right\rangle+\delta.$$

Taking expectations of both sides and rearranging gives

$$\mathbb{E}\left\{\left\langle Q_t\ ,\ x^* - x_{t+1}\right\rangle\right\} \leq \delta. \tag{20}$$

Lemma 4 *Lines 7, 8, and 9 of Algorithm 1 are equivalent to:*

$$y_{t+1} = \arg\min_{y\in\mathcal{Y}} \left\{\eta\left\langle Q_t\ ,\ y - x_{t+1}\right\rangle + \left\langle s_t\ ,\ y - y_t\right\rangle + \beta\left(W_t + h(y_t)\right)^\top l_t(y) + \frac{\eta}{2}\|y - x_{t+1}\|^2 + \frac{\alpha + 2G^2\beta}{2}\|y - y_t\|^2\right\}. \tag{21}$$

Proof See Appendix A. □

Lemma 5 *[Pushback lemma] Let function $r : \mathbb{V} \to \mathbb{R}$ be convex function and let $\mathcal{Y} \subseteq \mathbb{V}$ be a convex set. Fix $\zeta > 0$, $\tilde{x} \in \mathbb{V}$. Suppose there exists a point y such that:*

$$y = \arg\min_{x\in\mathcal{Y}} \left\{r(x) + \zeta\|x - \tilde{x}\|^2\right\}.$$

Then

$$r(y) + \zeta\|y - \tilde{x}\|^2 \leq r(z) + \zeta\|z - \tilde{x}\|^2 - \zeta\|z - y\|^2 \quad \forall z \in \mathcal{Y}.$$

Proof This lemma and its proof, which relies on the first-order optimality condition and the definition of strong convexity, can be found in various forms in [83–85]. □

Proof of Theorem 1 Fix $t \in \{1, 2, \ldots, T-1\}$. By definition of y_{t+1} as the minimizer in (21) we have by the pushback lemma (and the fact $x^* \in \mathcal{Y}$):

$$\begin{aligned}
&\eta\left\langle Q_t\ ,\ y_{t+1} - x_{t+1}\right\rangle + \left\langle s_t\ ,\ y_{t+1} - y_t\right\rangle + \beta\left(W_t + h(y_t)\right)^\top l_t(y_{t+1}) \\
&\quad + \frac{\eta}{2}\|y_{t+1} - x_{t+1}\|^2 + \frac{\alpha + 2G^2\beta}{2}\|y_{t+1} - y_t\|^2 \\
&\leq \eta\left\langle Q_t\ ,\ x^* - x_{t+1}\right\rangle + \left\langle s_t\ ,\ x^* - y_t\right\rangle + \beta\left(W_t + h(y_t)\right)^\top l_t(x^*) \\
&\quad + \frac{\eta}{2}\|x^* - x_{t+1}\|^2 + \frac{\alpha + 2G^2\beta}{2}\left(\|x^* - y_t\|^2 - \|x^* - y_{t+1}\|^2\right).
\end{aligned} \tag{22}$$

Denote the right-hand-side and left-hand-side of the inequality above as $\mathbf{RHS}_t$ and $\mathbf{LHS}_t$, respectively.

By completing the square, we obtain:

$$\left\langle s_t\ ,\ y_{t+1} - y_t\right\rangle + \frac{\alpha}{2}\|y_{t+1} - y_t\|^2 \geq -\frac{\|s_t\|^2}{2\alpha}.$$

Substituting this inequality into the $\mathbf{LHS}_t$ gives

$$\begin{aligned}\mathbf{LHS}_t \geq & \eta \langle Q_t\,,\ y_{t+1} - x_{t+1}\rangle + \beta\,(W_t + h(y_t))^\top l_t(y_{t+1}) - \frac{\|s_t\|^2}{2\alpha} \\ & + \frac{\eta}{2}\|y_{t+1} - x_{t+1}\|^2 + \frac{2G^2\beta}{2}\|y_{t+1} - y_t\|^2 \\ \geq & \frac{\eta}{2}\|Q_{t+1}\|^2 - \frac{\eta}{2}\|Q_t\|^2 + \frac{G^2\beta}{2}\|y_{t+1} - y_t\|^2 - \frac{\|s_t\|^2}{2\alpha} \\ & + \frac{\beta}{2}\|W_{t+1}\|_2^2 - \frac{\beta}{2}\|W_t\|_2^2 - \frac{\beta}{2}\|\left[-h(y_{t+1})\right]_+\|_2^2 + \frac{\beta}{2}\|h(y_t)\|_2^2\end{aligned}$$

where the last inequality uses Eqs. (12), and (18). By taking expectations and summing over $t \in \{1, 2, \ldots, T-1\}$, and using the inequality $\|[-x]_+\|_2 \leq \|x\|_2$, we obtain:

$$\begin{aligned}\sum_{t=1}^{T-1}\mathbb{E}\{\mathbf{LHS}_t\} \geq & \frac{\eta}{2}\mathbb{E}\left\{\|Q_T\|^2 - \|Q_1\|^2\right\} + \frac{G^2\beta}{2}\sum_{t=1}^{T-1}\mathbb{E}\left\{\|y_{t+1} - y_t\|^2\right\} \\ - \sum_{t=1}^{T-1}\frac{\mathbb{E}\left\{\|s_t\|^2\right\}}{2\alpha} & + \frac{\beta}{2}\mathbb{E}\left\{\|W_T\|_2^2 - \|W_1\|_2^2 - \|[-h(y_T)]_+\|_2^2 + \|h(y_1)\|_2^2\right\}\end{aligned} \tag{23}$$

Lines 2, 4, and 13 of Algorithm 1 lead to the following implications, respectively:

$$\begin{aligned}& Q_1 = \mathbf{0} \\ & \|W_1\|_2 = \|[-h(y_1)]_+\|_2 \leq \|h(y_1)\|_2 \\ & \|W_T\|_2 \geq \|[-h(y_T)]_+\|_2\end{aligned}$$

Utilizing the inequalities mentioned above and dropping the positive term $\|y_{t+1} - y_t\|^2$ (we will use $\|y_{t+1} - y_t\|^2$ when proving Theorem 2), Eq. (23) becomes:

$$\sum_{t=1}^{T-1}\mathbb{E}\{\mathbf{LHS}_t\} \geq \frac{\eta}{2}\mathbb{E}\left\{\|Q_T\|^2\right\} - \sum_{t=1}^{T-1}\frac{\mathbb{E}\left\{\|s_t\|^2\right\}}{2\alpha}. \tag{24}$$

Now consider the $\mathbf{RHS}_t$ of (22). Given Assumption 1, we have $\|x^* - x_{t+1}\| \leq D$. Using this and taking the expectation yields:

$$\begin{aligned}\mathbb{E}\{\mathbf{RHS}_t\} \leq & \eta\mathbb{E}\left\{\langle Q_t\,,\ x^* - x_{t+1}\rangle\right\} + \mathbb{E}\left\{\langle s_t\,,\ x^* - y_t\rangle\right\} \\ & + \beta\mathbb{E}\left\{(W_t + h(y_t))^\top l_t(x^*)\right\} + \frac{\eta D^2}{2} \\ & + \frac{\alpha + 2G^2\beta}{2}\mathbb{E}\left\{\|x^* - y_t\|^2 - \|x^* - y_{t+1}\|^2\right\}\end{aligned}$$

Using (17), which states that $\mathbb{E}\left\{(W_t + h(y_t))^\top l_t(x^*)\right\} \leq 0$, and (20), which states that $\mathbb{E}\{\langle Q_t\,,\, x^* - x_{t+1}\rangle\} \leq \delta$, we can further simplify the expression as follows:

$$\mathbb{E}\{\mathbf{RHS}_t\} \leq \eta\delta + \mathbb{E}\left\{\langle s_t\,,\, x^* - y_t\rangle\right\} + \frac{\alpha + 2G^2\beta}{2}\mathbb{E}\left\{\|x^* - y_t\|^2 - \|x^* - y_{t+1}\|^2\right\} + \frac{\eta D^2}{2}.$$

Line 1 of the algorithm states $y_1 = x_1 \in \mathcal{X}$ thus Assumption 1 implies $\|x^* - y_1\| \leq D$. Using this and summing over $t \in \{1, 2, \ldots, T-1\}$, we obtain:

$$\sum_{t=1}^{T-1}\mathbb{E}\{\mathbf{RHS}_t\} \leq \sum_{t=1}^{T-1}\mathbb{E}\left\{\langle s_t\,,\, x^* - y_t\rangle\right\} - \frac{\alpha + 2G^2\beta}{2}\mathbb{E}\left\{\|x^* - y_T\|^2\right\} + \frac{\alpha + 2G^2\beta}{2}D^2 + T\eta\frac{D^2 + 2\delta}{2}, \tag{25}$$

where in the last term we used $T - 1 \leq T$ to simplify it.

Substituting Eqs. (24) and (25) into (22) and rearranging the terms, we obtain:

$$\sum_{t=1}^{T-1}\mathbb{E}\left\{\langle s_t\,,\, y_t - x^*\rangle\right\} \leq -\frac{\eta}{2}\mathbb{E}\left\{\|Q_T\|^2\right\} - \frac{\alpha + 2G^2\beta}{2}\mathbb{E}\left\{\|x^* - y_T\|^2\right\} + \frac{\alpha + 2G^2\beta}{2}D^2 + T\eta\frac{D^2 + 2\delta}{2} + \sum_{t=1}^{T-1}\frac{\mathbb{E}\left\{\|s_t\|^2\right\}}{2\alpha} \tag{26}$$

Consider the following:

$$0 \leq \frac{1}{2}\left\|\frac{s_T}{\sqrt{\alpha}} + \sqrt{\alpha}(x^* - y_T)\right\|^2 = \frac{\|s_T\|^2}{2\alpha} + \frac{\alpha}{2}\|x^* - y_T\|^2 + \langle s_T\,,\, x^* - y_T\rangle$$

Taking expectation we can simply write

$$\mathbb{E}\left\{\langle s_T\,,\, y_T - x^*\rangle\right\} \leq \frac{\mathbb{E}\left\{\|s_T\|^2\right\}}{2\alpha} + \frac{\alpha}{2}\mathbb{E}\left\{\|x^* - y_T\|^2\right\} \tag{27}$$

Replacing (27) in (26) and dropping the negative term $-\frac{2G^2\beta}{2}\mathbb{E}\left\{\|x^* - y_T\|^2\right\}$, we get:

$$\sum_{t=1}^{T}\mathbb{E}\left\{\langle s_t\,,\, y_t - x^*\rangle\right\} \leq -\frac{\eta}{2}\mathbb{E}\left\{\|Q_T\|^2\right\} + \sum_{t=1}^{T}\frac{\mathbb{E}\left\{\|s_t\|^2\right\}}{2\alpha} + \frac{\alpha + 2G^2\beta}{2}D^2 + T\eta\frac{D^2 + 2\delta}{2} \tag{28}$$

Remember we defined $\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$. For the left-hand-side of (28) we can write:

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left\{\langle s_t\,,\; y_t - x^*\rangle\right\} &\overset{(a)}{\geq} \sum_{t=1}^{T}\left(\mathbb{E}\left\{f(y_t)\right\} - f(x^*)\right)\\ &\overset{(b)}{\geq} T\mathbb{E}\left\{f\left(\frac{1}{T}\sum_{t} = 1^T y_t\right)\right\} - Tf(x^*)\\ &= T\mathbb{E}\left\{f(\bar{y}_T)\right\} - Tf(x^*)\\ &\overset{(c)}{\geq} T\mathbb{E}\left\{f(\bar{x}_T) - L\|\bar{y}_T - \bar{x}_T\|\right\} - Tf(x^*).\end{aligned}$$

where (a) holds by Eq. (19); (b) holds by Jensen's inequality; and (c) relies on the Lipschitz continuity of f as established in Lemma 1. Substituting this in Eq. (28) we get:

$$\begin{aligned}T\mathbb{E}\left\{f(\bar{x}_T)\right\} - Tf(x^*) \leq TL\mathbb{E}\left\{\|\bar{y}_T - \bar{x}_T\|\right\} - \frac{\eta}{2}\mathbb{E}\left\{\|Q_T\|^2\right\}\\ + \sum_{t=1}^{T}\frac{\mathbb{E}\left\{\|s_t\|^2\right\}}{2\alpha} + \frac{\alpha + 2G^2\beta}{2}D^2 + T\eta\frac{D^2 + 2\delta}{2}.\end{aligned} \tag{29}$$

Equation (13) states $Q_T = T\,(\bar{y}_T - \bar{x}_T)$. Thus by completing the square we can write:

$$\begin{aligned}&TL\mathbb{E}\left\{\|\bar{y}_T - \bar{x}_T\|\right\} - \frac{\eta}{2}\mathbb{E}\left\{\|Q_T\|^2\right\}\\ &= TL\mathbb{E}\left\{\|\bar{y}_T - \bar{x}_T\|\right\} - \frac{\eta}{2}\mathbb{E}\left\{T^2\|\bar{y}_T - \bar{x}_T\|^2\right\} \leq \frac{L^2}{2\eta}.\end{aligned}$$

Finally, by employing the above inequality in (29) and dividing both sides by T, we obtain:

$$\mathbb{E}\left\{f(\bar{x}_T)\right\} - f(x^*) \leq \frac{L^2}{2T\eta} + \eta\frac{D^2 + 2\delta}{2} + \sum_{t=1}^{T}\frac{\mathbb{E}\left\{\|s_t\|^2\right\}}{2T\alpha} + \frac{\alpha D^2}{2T} + \frac{G^2D^2\beta}{T}$$

Using Assumption 3.iii to bound $\mathbb{E}\left\{\|s_t\|^2\right\} \leq L^2$ completes the proof. □

4 Numerical experiments

In this section, we experiment with our algorithm in two different scenarios. In Sect. 4.1, a regression problem is considered where the objective function is nonsmooth, and there are no functional constraints. Two versions of our algorithm, one using inexact LMO and the other using exact LMO, are compared with other existing approaches. In Sect. 4.2, we consider a small network flow problem and run our

algorithm on two different formulations of the same problem: the first one includes all the constraints in the feasible set $\mathcal{X}$ and has no functional constraints ($h \equiv 0$), while the second formulation removes some of the constraints from the set $\mathcal{X}$ and includes them as functional constraints. In the second case, the resulting set $\mathcal{X}$ has a simpler LMO, but at the cost of possible violations of those constraints formulated as functional constraints.

4.1 Robust reduced rank regression with nuclear norm relaxation

The problem of multi-output regression [86], which is a special case of multi-task learning [87], can be defined as follows. Given a dataset consisting of n samples, where each sample includes a response vector $y_i \in \mathbb{R}^q$ and a predictor vector $x_i \in \mathbb{R}^p$, we consider a multivariate linear regression model:[4]

$$y = cx + e.$$

Here, $y = (y_1, \ldots, y_n)$ is an $n \times q$ matrix of responses, and $x = (x_1, \ldots, x_n)$ is an $n \times p$ matrix of predictors. The c is a $q \times p$ coefficient matrix, and $e = (e_1, \ldots, e_n)$ is a $q \times n$ matrix of independently and identically distributed random errors. The **goal** is to estimate c.

In traditional linear regression, it's often assumed that the errors follow a Gaussian distribution, which works well when the data conforms to this assumption. However, in cases where the data contains outliers or exhibits heavy tails that deviate from the Gaussian distribution, this assumption does not hold.

To address this, we assume that the error matrix e in our model follows a Laplace distribution, also known as the double-exponential distribution. The Laplace distribution assigns more weight to the tails of the distribution compared to the Gaussian distribution, making it better suited for modeling data with outliers and heavy tails [77–79].

Reduced Rank Regression, introduced by Anderson in 1951 [88], is a specific form of multi-output regression. It operates under the assumption that the rank of the coefficient matrix c is small. This allows us to capture underlying patterns in the data efficiently. However, the low-rank constraint does not define a convex set. A convex relaxation of this constraint is possible by employing the nuclear norm function [89]. The nuclear norm encourages low-rank solutions by penalizing the sum of the singular values of c [90].[5]

[4] Note that, to comply with the common notation used in statistics literature, x and y are given information of the problem in this section and are not the two variables of our algorithm. Our algorithm will run to find $\bar{c}$ in this section.

[5] For a more detailed discussion about the nuclear norm and its characteristics, see Appendix B.

Thus, our optimization problem can be formulated as follows:

$$\begin{aligned} \text{Minimize:} \quad & f(c) := \frac{1}{n} \sum_{i=1}^{n} \|y_i - c x_i\|_2 \\ \text{Subject to:} \quad & c \in \mathcal{C}_\gamma := \{c \in \mathbb{R}^{q \times p} : \|c\|_* \leq \gamma\} \end{aligned} \tag{R4NR}$$

Here, $\gamma > 0$ is a hyperparameter, $\|\cdot\|_2$ denotes the l_2-norm of a column vector, and $\|\cdot\|_*$ denotes the nuclear norm of a matrix. The choice of the *nonsmooth* $\|\cdot\|_2$ loss function, as opposed to the *smooth* $\|\cdot\|_2^2$, which is common in regression, increases robustness against outliers. Intuitively, this approach is more robust as the loss grows linearly, rather than quadratically, when distancing from the true value [91, 92].

Numerical results We generated synthetic data using the following configuration: $n = 200$, $q = 300$, $p = 500$, and $\text{rank}(c) = 40$. Each element of the noise matrix, $\{e_{i,j}\}$, are i.i.d. samples of a Laplace distribution with parameters $\mu = 0$ and $b = 2$, and $\{x_{i,j}\}$ are i.i.d. samples of a standard normal distribution. We conducted simulations on four algorithms, three of which employed full SVD calculations:

- Our novel algorithm incorporating an exact LMO.
- P-MOLES, as detailed in [93].
- Projected Subgradient Descent (PGD).

Additionally, we implemented our algorithm using an inexact LMO. [6] The figures are derived from the average of ten distinct runs of the experiment, ensuring statistical reliability.

Figures 1 and 2 illustrate the expected loss relative to noiseless data for a fixed γ value of 350. The x-axes of the respective figures represent the number of iterations and the computational time. The computational gain obtained by using inexact LMO is quite visible in Fig. 2.

In Fig. 3, we keep the number of iterations fixed at $T = 300$ and run the algorithm for different values of γ. When γ is very small, all four algorithms perform poorly (underfitting). On the other hand, as γ becomes very large, the performance of all models starts to deteriorate (overfitting). It is worth noting that the PGD algorithm underperforms in comparison to our projection-free algorithm and the P-MOLES algorithm. This can be explained by the fact that, as mentioned before, methods like Frank–Wolfe implicitly encourage sparsity. [7]

4.2 Minimum-cost flow in a convex-cost network

Any network optimization problem with general convex costs and side constraints can be formulated into the form of Problem (P1) (see, for example, [74]). There are

[6] See Appendix C for a more detailed discussion.

[7] This implicit regularization effect diminishes as the number of iterations increases. It remains an open question how to disentangle the number of iterations and the degree of sparsity enforced by Frank–Wolfe-type algorithms. One possible idea is to restart the algorithm after some iterations using the previous point as the initial point.

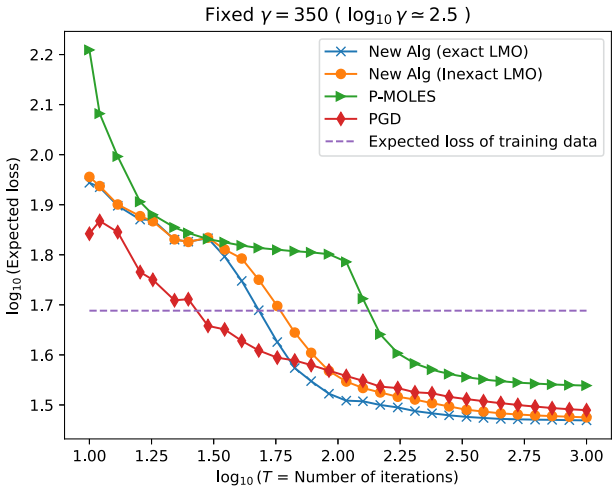

Fig. 1 Expected loss as a function of the number of iterations for $\gamma = 350$, compared to noiseless data.

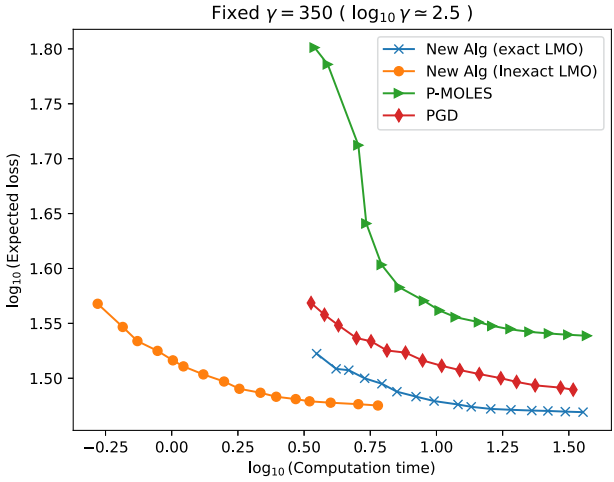

Fig. 2 Computational time versus error for inexact LMO with $\gamma = 350$, demonstrating the computational efficiency.

numerous high-performance specialized algorithms available for linear network optimization problems. This is noteworthy because our approach effectively addresses a nonlinear optimization problem by sequentially solving multiple linear optimizations. The original Frank–Wolfe algorithm was used for such problems where there are no functional constraints and the cost function is smooth [94, 95].

Consider a directed acyclic graph with a vertex set V and an edge set E. Each edge $(i, j) \in E$ has an associated nonnegative capacity k_{ij}, which represents the maximum allowable flow over that edge. Let $x_{ij} \in \mathbb{R}_+$ denote the flow over edge (i, j). The cost of routing a flow vector $x \in \mathbb{R}_+^{|E|}$ through this network is given by a cost function

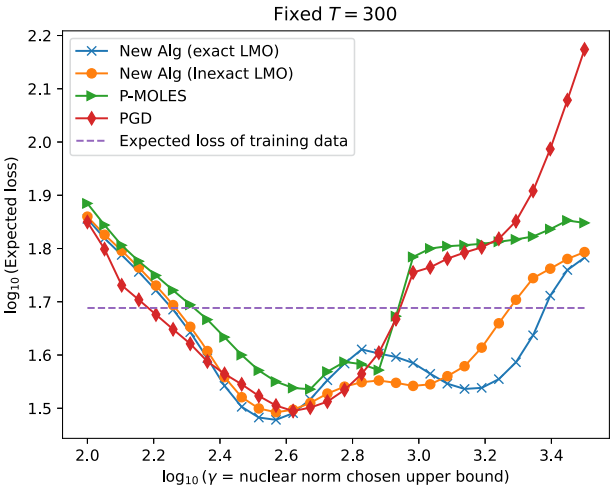

Fig. 3 Performance of four algorithms at $T = 300$ iterations. Performance dips for very small or large γ

$f(x)$. Our objective is to route a fixed flow amount $d > 0$ from a source node s to a target node t at the minimum possible cost. Formally, the problem can be defined as follows:

$$\begin{aligned} \text{Minimize:} \quad & f(x) := \sum_{(i,j)\in E} \max\left\{a_{ij}x_{ij} + b_{ij}\,,\; c_{ij}\right\} \qquad \text{(Min-Flow)} \\ \text{Subject to:} \quad & x \in \mathcal{P} \\ & x_{ij} \le k_{ij}, \quad \forall ij \in E \end{aligned}$$

The known nonnegative constants a_{ij}, b_{ij}, and c_{ij} describe the nonsmooth Lipschitz continuous convex function $f : \mathbb{R}^{|E|} \to \mathbb{R}$. The set $\mathcal{P}$ comprises all flows that satisfy the flow conservation laws, without considering the capacities, and is defined as:

$$\mathcal{P} := \left\{ x \in \mathbb{R}_+^{|E|} : \sum_{\{j:(i,j)\in E\}} x_{ij} - \sum_{\{j:(j,i)\in E\}} x_{ji} = r_i,\ \forall i \in V \right\}$$

In our example, the vector $r \in \mathbb{R}^{|V|}$ is given by:[8]

$$r_i = \begin{cases} -d & \text{if } i = t, \\ +d & \text{if } i = s, \\ 0 & \text{otherwise.} \end{cases}$$

[8] The vector r represents the net demand or supply at each node in the network.

Table 1 Overview of four reformulations of Problem (Min-Flow) in the general form of Problem (P2)

	Formulation 1	Formulation 2	Formulation 3	Formulation 4
$\mathcal{X}$	$\mathcal{X}_{\text{CAP}}$	$\mathcal{X}_{\text{CAP}}$	$\mathcal{P}$	$\mathcal{P}$
$\mathcal{Y}$	$\mathbb{R}^{\lvert E\rvert}$	$\mathcal{Y}_{\text{BOX}}$	$\mathbb{R}^{\lvert E\rvert}$	$\mathcal{Y}_{\text{BOX}}$
h	–	–	h_{CAP}	h_{CAP}

We reframe Problem (Min-Flow) into the structure of (P2) in four distinct ways, as detailed in Table 1. Define the following two sets:

$$\mathcal{X}_{\text{CAP}} := \mathcal{P} \cap \left\{ x \in \mathbb{R}_+^{|E|} : x_{ij} \leq k_{ij}, \forall (i,j) \in E \right\},$$
$$\mathcal{Y}_{\text{BOX}} := \left\{ x \in \mathbb{R}_+^{|E|} : x_{ij} \leq \max\{d, k_{ij}\}, \forall (i,j) \in E \right\},$$

and the following nonsmooth, Lipschitz continuous convex constraint function:

$$h_{\text{CAP}}(x) := \max \left\{ x_{ij} - k_{ij} : (i,j) \in E \right\}.$$

Choosing $\mathcal{X} = \mathcal{X}_{\text{CAP}}$ eliminates the need for functional constraints, while choosing $\mathcal{X} = \mathcal{P}$ enforces the capacity constraints using $h(x) \leq 0$. This has the following consequences:

1. The linear minimization on the set $\mathcal{P}$ corresponds to a *shortest path* problem [74], whereas the linear minimization on $\mathcal{X}_{\text{CAP}}$ represents a minimum-cost flow problem with a linear cost function. In terms of implementation, finding the shortest path on a directed acyclic graph can be achieved in $\Theta(|V|+|E|)$ using *topological sorting* [96]. In contrast, solving a linear minimum-cost flow problem is challenging; however, the *network simplex algorithm* in practice operates with an average time complexity of $\mathcal{O}(|E| \cdot |V|)$ [97].
2. This computational gain may come at the cost of violating capacity constraints, i.e., $h(\bar{x}) \not\leq 0$.

The two different choices of $\mathcal{Y} = \mathcal{Y}_{\text{BOX}}$ and $\mathcal{Y} = \mathbb{R}^{|E|}$ are made to observe if choosing a smaller set $\mathcal{Y}$ affects our algorithm's performance. Both of these sets satisfy our assumptions of easy projection and $\mathcal{X} \subseteq \mathcal{Y}$.

The network graph that we used is shown in Fig. 4. The numbers on the edges indicate the capacities, k_{ij}. The parameters were set as follows: $d = 4.1$, $a_{ij} = \exp\left(k_{ij}/10\right)$, $b_{ij} = k_{ij}/10$, and $c_{ij} = k_{ij}/5$. Figure 5 illustrates the cost versus the number of iterations parameter T on a log-log scale for all four formulations. No significant difference between the two choices of the set $\mathcal{Y}$ is visible. While Formulations 3 and 4 converge slower than Formulations 1 and 2 for the same number of iterations T, they compensate in actual running time with a faster LMO.[9] Figure 6

[9] The computational gain difference between linear minimization on $\mathcal{X}_{\text{CAP}}$ and $\mathcal{P}$ was significant in our experiment, approximately 50-fold. However, the corresponding figure is not depicted here, as we did not optimize either LMO.

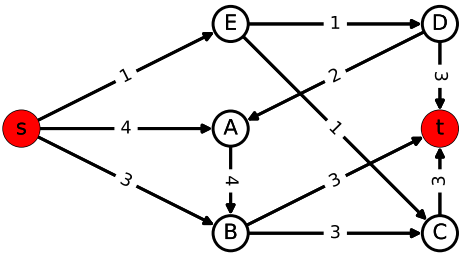

Fig. 4 The directed acyclic graph used in the experiment has nonnegative capacities on each edge. The objective is to route a fixed amount of flow from the source node s to the target node t at the minimum possible cost

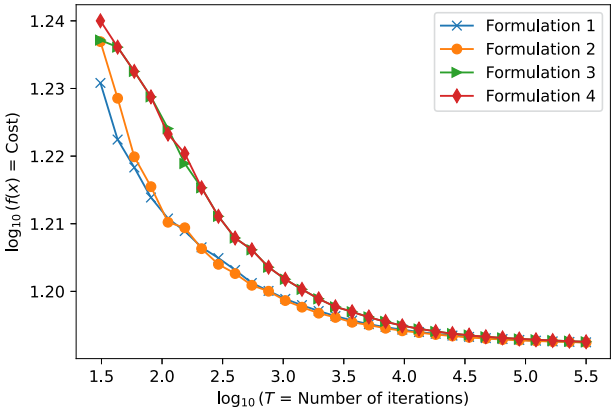

Fig. 5 The cost vs. number of iterations parameter T is shown on a log-log scale for all four formulations. There is no significant difference between the two $\mathcal{Y}$ choices. While Formulations 3 and 4 converge slower than 1 and 2, they compensate with a faster LMO oracle (not depicted here)

shows the value of the functional constraint $h(x)$ (positive values indicate violation). This plot highlights that our algorithm's output may violate the capacity constraints when enforced by functional constraints (as in Formulations 3 and 4) instead of set constraints (as in Formulations 1 and 2).

5 Conclusions and open problems

This paper tackles the problem of solving general convex optimization with functional constraints without projecting onto the feasible set. Previous studies on projection-free algorithms mainly focused on smooth problems and/or did not consider functional constraints. Our experiments and convergence theorems demonstrate that our algorithm performs comparably to projected stochastic subgradient descent methods, making it a viable alternative in scenarios where projection-free approaches are preferred.

An open problem is whether our algorithm can incorporate benefits from mirror descent [98, 99], an established method that substitutes the norm in projected subgradient descent with Bregman divergence, leading to enhanced performance in specific contexts, such as when dealing with a probability simplex. Another question is whether, similar to projected subgradient descent, we can achieve an improved rate for nonsmooth, *strongly convex* optimization [100]. Another important area to explore is how

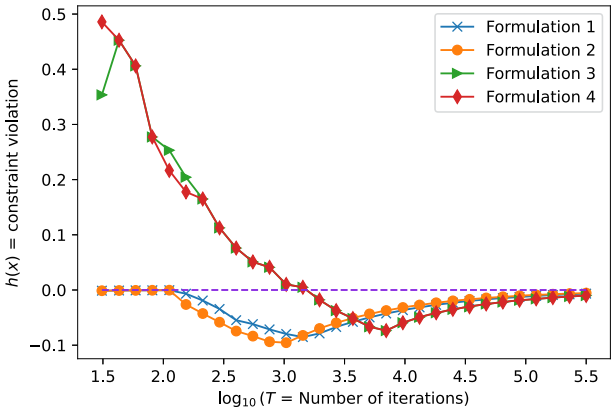

Fig. 6 The figure illustrates the value of the constraint function $h(x)$ for all four formulations. Values below zero indicate no violation of capacity constraints. Formulations 1 and 2 consistently stay within the capacity region, so no violations occur for them

our method works in online settings. This is especially relevant since projection-free online convex optimization is a highly discussed topic today [37, 60, 61, 64, 65, 101].

Additionally, it is worth investigating whether the subgradient can be computed on points within the feasible set $\mathcal{X}$, rather than using $y_t \in \mathcal{Y}$ that may lie outside of the set $\mathcal{X}$. Notably, our algorithm is not unique in utilizing subgradients outside the feasible set [36, 38, 62, 66, 67, 93, 102].

Funding Open access funding provided by SCELC, Statewide California Electronic Library Consortium This research received partial support from the National Science Foundation (NSF), including NSF CCF 1718477 and NSF SpecEES 1824418 grants.

Data availability The data used in our simulation is synthetic.

Code Availability The codes and synthetically generated data are accessible at github.com/kamiarasgari.

Declarations

Conflict of interest The authors declare that they have no competing interests relevant to the content of this article.

Appendix A Remaining proofs

Proof of Lemma 1 **Lipschitz continuity of** f: Let $x, y \in \mathcal{Y}$. Consider stochastic subgradients s_x of f at x. This vector, as per Assumption 3.iii, satisfies the following conditions:

$$\mathbb{E}\left\{s_x \middle| x\right\} \in \partial f(x), \quad \sqrt{\mathbb{E}\left\{\|s_x\|^2 \middle| x\right\}} \leq L.$$

Exploiting the convexity of f and applying the Cauchy-Schwarz inequality, we arrive at:

$$f(x) - f(y) \leq \left\langle \mathbb{E}\{s_x \middle| x\}, x - y \right\rangle \leq \left\| \mathbb{E}\{s_x \middle| x\} \right\| \|x - y\|.$$

Using Jensen's inequality for the norm function (which is convex) and the nonnegativity of the variance, we can further deduce:

$$\left\| \mathbb{E}\{s_x \middle| x\} \right\| \leq \mathbb{E}\left\{\|s_x\| \middle| x\right\} \leq \sqrt{\mathbb{E}\left\{\|s_x\|^2 \middle| x\right\}} \leq L$$

These implications lead to:

$$f(x) - f(y) \leq L\|x - y\|.$$

Similarly, considering a stochastic subgradient s_y of f at point y, we can deduce:

$$f(y) - f(x) \leq L\|x - y\|.$$

This concludes the proof of Lipschitz continuity for f.

Lipschitz continuity of h_i: Let $x, y \in \mathcal{Y}$. Consider stochastic subgradients g_x of h_i at x. This vector, as per Assumption 3.iii, satisfy the following conditions:

$$\mathbb{E}\left\{g_x \middle| x\right\} \in \partial h_i(x), \quad \|g_x\| \leq G_i.$$

Similar to the analysis of function f, by exploiting the convexity of h_i, applying the Cauchy-Schwarz inequality, and using Jensen's inequality we arrive at:

$$h_i(x) - h_i(y) \leq \mathbb{E}\left\{\|g_x\| \middle| x\right\} \|x - y\| \leq G_i\|x - y\|.$$

Similarly, considering a stochastic subgradient g_y of h_i at point y, we can deduce:

$$h_i(y) - h_i(x) \leq G_i\|x - y\|.$$

This concludes the proof of Lipschitz continuity for h_i.

Lipschitz continuity of h: Lipschitz continuity of h_i implies:

$$(h_i(y) - h_i(x))^2 \leq G_i^2\|x - y\|^2.$$

By summing over $i \in \{1, \ldots, m\}$ we get:

$$\|h(y) - h(x)\|_2^2 \leq \sum_{i=1}^{m} G_i^2 \|x - y\|^2 \leq G^2 \|x - y\|^2.$$

This concludes the proof of Lipschitz continuity for h. □

Proof of Lemma 4 We initiate our proof with Line 9, which states that $y_{t+1} = \mathrm{PO}_{\mathcal{Y}}\{\tilde{y}_{t+1}\}$. By applying the definition of projection from Eq. (2), we can express it as:

$$y_{t+1} = \arg\min_{y \in \mathcal{Y}} \{\|y - \tilde{y}_{t+1}\|\} = \arg\min_{y \in \mathcal{Y}} \{\|y - \tilde{y}_{t+1}\|^2\}.$$

Now, utilizing Line 8, we obtain:

$$\|y - \tilde{y}_{t+1}\|^2 = \left\| y - \frac{\left(\alpha + 2G^2\beta\right) y_t + \eta x_{t+1} - p_t}{\alpha + 2G^2\beta + \eta} \right\|^2$$

Define $\Omega := \alpha + 2G^2\beta + \eta$. This allows us to further simplify it as:

$$y_{t+1} = \arg\min_{y \in \mathcal{Y}} \left\{ \Omega \|y\|^2 - 2\left\langle \left(\alpha + 2G^2\beta\right) y_t + \eta x_{t+1} \,,\, y \right\rangle + 2\langle p_t \,,\, y \rangle \right\}$$

Continuing with the simplification and invoking Line 7, which defines the temporary variable p_t, we arrive at:

$$\begin{aligned} y_{t+1} = &\arg\min_{y \in \mathcal{Y}} \left\{ \frac{\eta}{2} \|y - x_{t+1}\|^2 + \frac{\alpha + 2G^2\beta}{2} \|y - y_t\|^2 \right. \\ &\left. + \langle \eta Q_t \,,\, y \rangle + \langle s_t \,,\, y \rangle + \left\langle \beta \sum_{i=1}^{m} (W_{i,t} + h_i(y_t)) g_{i,t} \,,\, y \right\rangle \right\} \\ = &\arg\min_{y \in \mathcal{Y}} \left\{ \frac{\eta}{2} \|y - x_{t+1}\|^2 + \frac{\alpha + 2G^2\beta}{2} \|y - y_t\|^2 + \eta \langle Q_t \,,\, y - x_{t+1} \rangle \right. \\ &\left. + \langle s_t \,,\, y - y_t \rangle + \beta \sum_{i=1}^{m} (W_{i,t} + h_i(y_t)) \left(h_i(y_t) + \langle g_{i,t} \,,\, y \rangle \right) \right\}. \end{aligned}$$

Finally, by utilizing the linearized function l_t defined in Eq. (14), we conclude the proof. □

A.1 Proof of Theorem 2

Lemma 6 *Line 13 implies the following inequality:*

$$\| [h(\bar{x}_T)]_+ \|_2 \leq \frac{\|W_T\|_2 + \| [h(y_T)]_+ \|_2}{T} + \frac{G}{T} \sum_{t=1}^{T-1} \|y_{t+1} - y_t\| + G\|\bar{y}_T - \bar{x}_T\|.$$

Proof Fix $i \in \{1, \ldots, m\}$. Line 13 of the algorithm implies:

$$W_{i,t+1} \geq W_{i,t} + h_i(y_t) + \langle g_{i,t}\,,\ y_{t+1} - y_t \rangle$$

Using the Cauchy-Schwarz inequality we get:

$$W_{i,t+1} \geq W_{i,t} + h_i(y_t) - \|g_{i,t}\|\,\|y_{t+1} - y_t\|$$

Summing over $t \in \{1, \ldots, T-1\}$ gives:

$$W_{i,T} \geq W_{i,1} + \sum_{t=1}^{T-1} h_i(y_t) - \sum_{t=1}^{T-1} \|g_{i,t}\|\,\|y_{t+1} - y_t\|$$

Using the fact that $W_{i,1} \geq 0$ from Lemma 3, and adding $h_i(y_T)$ to both sides, we get:

$$\sum_{t=1}^{T} h_i(y_t) \leq W_{i,T} + h_i(y_T) + \sum_{t=1}^{T-1} \|g_{i,t}\|\,\|y_{t+1} - y_t\|$$

Furthermore, by applying Jensen's inequality we get:

$$h_i(\bar{y}_T) \leq \frac{1}{T}\left(W_{i,T} + h_i(y_T) + \sum_{t=1}^{T-1} \|g_{i,t}\|\,\|y_{t+1} - y_t\|\right)$$

Adding $h_i(\bar{x}_T) - h_i(\bar{y}_T)$ to both sides of the inequality we get

$$h_i(\bar{x}_T) \leq \frac{1}{T}\left(W_{i,T} + h_i(y_T) + \sum_{t=1}^{T-1} \|g_{i,t}\|\,\|y_{t+1} - y_t\|\right) + h_i(\bar{x}_T) - h_i(\bar{y}_T)$$

The positive part function $[\cdot]_+$ is nondecreasing. Therefore:

$$[h_i(\bar{x}_T)]_+ \leq \left[\frac{1}{T}\left(W_{i,T} + h_i(y_T) + \sum_{t=1}^{T-1} \|g_{i,t}\|\,\|y_{t+1} - y_t\|\right) + h_i(\bar{x}_T) - h_i(\bar{y}_T)\right]_+$$

Lemma 3, states $W_{i,T} \geq 0$. Using the general property that for any two nonnegative real numbers a and b, $[a+b]_+ \leq [a]_+ + [b]_+$, we obtain:

$$[h_i(\bar{x}_T)]_+ \leq \frac{W_{i,T} + \big[h_i(y_T)\big]_+}{T} + \frac{1}{T}\sum_{t=1}^{T-1} \|g_{i,t}\| \, \|y_{t+1} - y_t\| + \big[h_i(\bar{x}_T) - h_i(\bar{y}_T)\big]_+ \tag{A1}$$

To continue the proof, consider the following inequality. Fix arbitrary vectors $a, b_1, \ldots, b_K \in \mathbb{R}^m_+$. If vector a is component-wise smaller than or equal to the vector $\sum_{k=1}^K b_k$, then we have $\|a\|_2 \leq \left\|\sum_{k=1}^K b_k\right\|_2$. Utilizing the triangle inequality this gives: $\|a\|_2 \leq \sum_{k=1}^K \|b_k\|_2$. Thus, considering (A1) as an inequality for i-th element of vectors belonging to $\mathbb{R}^m_+$, we can write: [10]

$$\begin{aligned}
\|[h(\bar{x}_T)]_+\|_2 \leq & \frac{\|W_T\|_2 + \left\|\big[h_i(y_T)\big]_+\right\|_2}{T} + \frac{1}{T}\sum_{t=1}^{T-1}\sqrt{\sum_{i=1}^m \|g_{i,t}\|^2}\,\|y_{t+1} - y_t\| \\
& + \left\|[h(\bar{x}_T) - h(\bar{y}_T)]_+\right\|_2 \\
\overset{(a)}{\leq} & \frac{\|W_T\|_2 + \left\|\big[h_i(y_T)\big]_+\right\|_2}{T} + \frac{G}{T}\sum_{t=1}^{T-1}\|y_{t+1} - y_t\| + \left\|[h(\bar{x}_T) - h(\bar{y}_T)]_+\right\|_2 \\
\overset{(b)}{\leq} & \frac{\|W_T\|_2 + \left\|\big[h_i(y_T)\big]_+\right\|_2}{T} + \frac{G}{T}\sum_{t=1}^{T-1}\|y_{t+1} - y_t\| + \|h(\bar{x}_T) - h(\bar{y}_T)\|_2 \\
\overset{(c)}{\leq} & \frac{\|W_T\|_2 + \left\|\big[h_i(y_T)\big]_+\right\|_2}{T} + \frac{G}{T}\sum_{t=1}^{T-1}\|y_{t+1} - y_t\| + G\|\bar{y}_T - \bar{x}_T\|
\end{aligned}$$

Here, (a) is by Assumption 3.iii; the simple fact that for any $a \in \mathbb{R}^m$, we have $\|[x]_+\|_2 \leq \|x\|_2$ implies (b); and (c) is by Lemma 1. □

Proof of Theorem 2 The initial steps of this proof closely resemble those in the proof of Theorem 1. Just as we did in that proof, we will denote the right-hand-side and left-hand-side of (22) as $\mathbf{RHS}_t$ and $\mathbf{LHS}_t$, respectively. The previously derived Eq. (23) is

[10] The vectors are as follows:

- $([h_1(\bar{x}_T)]_+, \ldots, [h_m(\bar{x}_T)]_+)^\top$,
- $\frac{1}{T}\left(W_{1,T}, \ldots, W_{m,T}\right)^\top$,
- $\frac{1}{T}\left([h_1(y_T)]_+, \ldots, [h_m(y_T)]_+\right)^\top$,
- $\frac{1}{T}\left(\|g_{1,t}\|\,\|y_{t+1} - y_t\|, \ldots, \|g_{m,t}\|\,\|y_{t+1} - y_t\|\right)^\top$, for all $t \in \{1, \ldots, T-1\}$,
- $\left([h_1(\bar{x}_T) - h_1(\bar{y}_T)]_+, \ldots, [h_m(\bar{x}_T) - h_m(\bar{y}_T)]_+\right)^\top$.

demonstrated here:

$$\sum_{t=1}^{T-1}\mathbb{E}\left\{\mathbf{LHS}_t\right\} \geq \frac{\eta}{2}\mathbb{E}\left\{\|Q_T\|^2 - \|Q_1\|^2\right\} + \frac{G^2\beta}{2}\sum_{t=1}^{T-1}\mathbb{E}\left\{\|y_{t+1}-y_t\|^2\right\} - \sum_{t=1}^{T-1}\frac{\mathbb{E}\left\{\|s_t\|^2\right\}}{2\alpha} + \frac{\beta}{2}\mathbb{E}\left\{\|W_T\|_2^2 - \|W_1\|_2^2 - \|\left[-h(y_T)\right]_+\|_2^2 + \|h(y_1)\|_2^2\right\} \quad \text{(Eq.(23) copied)}$$

Lines 2 and 4 of the algorithm lead to the following implications, respectively:

$$Q_1 = \mathbf{0},$$
$$\|W_1\|_2 = \|\left[-h(y_1)\right]_+\|_2 \leq \|h(y_1)\|_2.$$

Utilizing the inequalities mentioned above, we obtain:

$$\sum_{t=1}^{T-1}\mathbb{E}\left\{\mathbf{LHS}_t\right\} \geq \frac{\eta}{2}\mathbb{E}\left\{\|Q_T\|^2\right\} + \frac{G^2\beta}{2}\sum_{t=1}^{T-1}\mathbb{E}\left\{\|y_{t+1}-y_t\|^2\right\} - \sum_{t=1}^{T-1}\frac{\mathbb{E}\left\{\|s_t\|^2\right\}}{2\alpha} + \frac{\beta}{2}\mathbb{E}\left\{\|W_T\|_2^2 - \|\left[-h(y_T)\right]_+\|_2^2\right\} \quad \text{(A2)}$$

Note that, unlike what we did in (24), we did not utilize the inequality $\|W_T\|_2 \geq \|\left[-h(y_T)\right]_+\|_2$ in (A2).

For the $\mathbf{RHS}_t$, we use the previously derived (25), which is demonstrated below:

$$\sum_{t=1}^{T-1}\mathbb{E}\left\{\mathbf{RHS}_t\right\} \leq \sum_{t=1}^{T-1}\mathbb{E}\left\{\langle s_t\,,\, x^* - y_t\rangle\right\} - \frac{\alpha+2G^2\beta}{2}\mathbb{E}\left\{\|x^*-y_T\|^2\right\} + \frac{\alpha+2G^2\beta}{2}D^2 + T\eta\frac{D^2+2\delta}{2}. \quad \text{(Eq.(25) copied)}$$

Using (27) in the above equation, we get:

$$\sum_{t=1}^{T-1}\mathbb{E}\left\{\mathbf{RHS}_t\right\} \leq \sum_{t=1}^{T}\mathbb{E}\left\{\langle s_t\,,\, x^* - y_t\rangle\right\} - G^2\beta\mathbb{E}\left\{\|x^*-y_T\|^2\right\} + \frac{\alpha+2G^2\beta}{2}D^2 + T\eta\frac{D^2+2\delta}{2} + \frac{\mathbb{E}\left\{\|s_T\|^2\right\}}{2\alpha}. \quad \text{(A3)}$$

Consider the following derivation. Starting from Eq. (19), we obtain the following expression:

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left\{\langle s_t\,,\,x^*-y_t\rangle\right\} &\leq \sum_{t=1}^{T}\mathbb{E}\left\{f(x^*)-f(y_t)\right\}\\ \text{(by Lemma 2)} &\leq \sum_{t=1}^{T}\mathbb{E}\left\{\mu^\top h(y_t)+\langle\lambda\,,\,x_t-y_t\rangle\right\}\\ \text{(by the definition of } \bar{x}_T,\bar{y}_T) &= \sum_{t=1}^{T}\mathbb{E}\left\{\mu^\top h(y_t)\right\}+T\,\langle\lambda\,,\,\mathbb{E}\{\bar{x}_T-\bar{y}_T\}\rangle\end{aligned} \tag{A4}$$

For any $i\in\{1,\ldots,m\}$, Line 13 of the algorithm implies:

$$\begin{aligned}h_i(y_t) &\leq W_{i,t+1}-W_{i,t}-\langle g_{i,t}\,,\,y_{t+1}-y_t\rangle\\ \text{(by Cauchy-Schwarz)} &\leq W_{i,t+1}-W_{i,t}+\|g_{i,t}\|\,\|y_{t+1}-y_t\|\end{aligned} \tag{A5}$$

Summing (A5) over t in the range $t\in\{1,\ldots,T-1\}$ yields:

$$\begin{aligned}\sum_{t=1}^{T-1}h_i(y_t) &\leq \sum_{t=1}^{T-1}\left(W_{i,t+1}-W_{i,t}+\|g_{i,t}\|\,\|y_{t+1}-y_t\|\right)\\ &= W_{i,T}-W_{i,1}+\sum_{t=1}^{T-1}\|g_{i,t}\|\,\|y_{t+1}-y_t\|\\ (W_{i,1}\geq 0 \text{ by Lemma 3}) &\leq W_{i,T}+\sum_{t=1}^{T-1}\|g_{i,t}\|\,\|y_{t+1}-y_t\|\end{aligned} \tag{A6}$$

Multiplying both sides of (A6) by $\mu_i\geq 0$ and summing over i yields:

$$\begin{aligned}\sum_{i=1}^{m}\mu_i\sum_{t=1}^{T-1}h_i(y_t) &\leq \mu^\top W_T+\sum_{i=1}^{m}\mu_i\sum_{t=1}^{T-1}\|g_{i,t}\|\,\|y_{t+1}-y_t\|\\ &= \mu^\top W_T+\sum_{t=1}^{T-1}\left(\sum_{i=1}^{m}\mu_i\|g_{i,t}\|\right)\|y_{t+1}-y_t\|\\ \text{(by Cauchy-Schwarz)} &\leq \mu^\top W_T+\sum_{t=1}^{T-1}\left(\|\mu\|_2\sqrt{\sum_{i=1}^{m}\|g_{i,t}\|^2}\right)\|y_{t+1}-y_t\|\\ \text{(by Assumption 3.iii)} &\leq \mu^\top W_T+G\|\mu\|_2\sum_{t=1}^{T-1}\|y_{t+1}-y_t\|\end{aligned}$$

This gives us:

$$\sum_{t=1}^{T} \mu^\top h(y_t) = \mu^\top h(y_T) + \sum_{i=1}^{m} \mu_i \sum_{t=1}^{T-1} h_i(y_t)$$
$$\leq \mu^\top h(y_T) + \mu^\top W_T + G\|\mu\|_2 \sum_{t=1}^{T-1} \|y_{t+1} - y_t\|$$

Plugging the above inequality into (A4) results in:

$$\sum_{t=1}^{T} \mathbb{E}\left\{\langle s_t\,,\, x^* - y_t\rangle\right\} \leq \mathbb{E}\left\{\mu^\top h(y_T) + \mu^\top W_T\right\}$$
$$+ G\|\mu\|_2 \sum_{t=1}^{T-1} \mathbb{E}\left\{\|y_{t+1} - y_t\|\right\} + T\,\langle \lambda\,,\, \mathbb{E}\left\{\bar{x}_T - \bar{y}_T\right\}\rangle \tag{A7}$$

Substituting Eqs. (A2), (A3), and (A7) into (22) and rearranging the terms, we obtain:

$$\begin{aligned}
&\frac{\eta}{2}\mathbb{E}\left\{\|Q_T\|^2\right\} - T\,\langle \lambda\,,\, \mathbb{E}\left\{\bar{x}_T - \bar{y}_T\right\}\rangle \\
&+ \frac{G^2\beta}{2}\sum_{t=1}^{T-1}\mathbb{E}\left\{\|y_{t+1} - y_t\|^2\right\} - G\|\mu\|_2 \sum_{t=1}^{T-1}\mathbb{E}\left\{\|y_{t+1} - y_t\|\right\} \\
&+ \frac{\beta}{2}\mathbb{E}\left\{\|W_T\|_2^2 - \|\,[-h(y_T)]_+\,\|_2^2\right\} - \mathbb{E}\left\{\mu^\top h(y_T) + \mu^\top W_T\right\} \\
&\leq \sum_{t=1}^{T}\frac{\mathbb{E}\left\{\|s_t\|^2\right\}}{2\alpha} - G^2\beta\mathbb{E}\left\{\|x^* - y_T\|^2\right\} + \frac{\alpha + 2G^2\beta}{2}D^2 + T\eta\frac{D^2 + 2\delta}{2}
\end{aligned} \tag{A8}$$

Next, we simplify the following two terms from the above equation: $\frac{\beta}{2}\mathbb{E}\left\{-\|\,[-h(y_T)]_+\,\|_2^2\right\}$ from the left-hand side and $-G^2\beta\mathbb{E}\left\{\|x^* - y_T\|^2\right\}$ from the right-hand side. Using the Lipschitz continuity of h (see Lemma 1), we get:

$$\left\|h(y_T) - h(x^*)\right\|_2 \leq G\left\|x^* - y_T\right\|$$

By using reverse triangle inequality we get

$$\|h(y_T)\|_2 \leq \left\|h(x^*)\right\|_2 + G\left\|x^* - y_T\right\|$$

Using the simple fact that $(a+b)^2 \leq 2a^2 + 2b^2$ we get:

$$\|h(y_T)\|_2^2 \leq 2\left\|h(x^*)\right\|_2^2 + 2G^2\left\|x^* - y_T\right\|^2$$

Using the fact that for any arbitrary vector $v \in \mathbb{R}^m$, the inequality $\|v\|_2^2 = \|[-v]_+\|_2^2 + \|[v]_+\|_2^2$ holds, we can write:

$$\|[-h(y_T)]_+\|_2^2 \leq 2\|h(x^*)\|^2 + 2G^2\|x^* - y_T\|^2 - \|[h(y_T)]_+\|_2^2 \tag{A9}$$

Utilizing Eq. (A9) within (A8) and further simplifying by applying $Q_T = T(\bar{y}_T - \bar{x}_T)$, and the inequality $\mathbb{E}\{\|s_t\|^2\} \leq L^2$, we obtain:

$$\begin{aligned}
(\mathbf{LHS}_a:=) \quad & \frac{T^2\eta}{2}\mathbb{E}\left\{\|\bar{y}_T - \bar{x}_T\|^2\right\} - T\left\langle \lambda\,,\, \mathbb{E}\{\bar{x}_T - \bar{y}_T\}\right\rangle \\
(\mathbf{LHS}_b:=) \quad & + \frac{G^2\beta}{2}\sum_{t=1}^{T-1}\mathbb{E}\left\{\|y_{t+1} - y_t\|^2\right\} - G\|\mu\|_2\sum_{t=1}^{T-1}\mathbb{E}\{\|y_{t+1} - y_t\|\} \\
(\mathbf{LHS}_c:=) \quad & + \frac{\beta}{2}\mathbb{E}\left\{\|W_T\|_2^2 + \|[h(y_T)]_+\|_2^2\right\} - \mathbb{E}\left\{\mu^\top h(y_T) + \mu^\top W_T\right\} \\
& \leq \beta\|h(x^*)\|_2^2 + \frac{TL^2}{2\alpha} + \frac{\alpha + 2G^2\beta}{2}D^2 + T\eta\frac{D^2 + 2\delta}{2}
\end{aligned} \tag{A10}$$

Here, the left-hand-side is divided into three terms, each of which is simplified as follows:

$$\begin{aligned}
\mathbf{LHS}_a =& \frac{T^2\eta}{2}\mathbb{E}\left\{\|\bar{y}_T - \bar{x}_T\|^2\right\} - T\left\langle \lambda\,,\, \mathbb{E}\{\bar{x}_T - \bar{y}_T\}\right\rangle \\
\overset{(a)}{\geq}& \frac{T^2\eta}{2}\left(\mathbb{E}\{\|\bar{y}_T - \bar{x}_T\|\}\right)^2 - T\|\lambda\|\mathbb{E}\{\|\bar{y}_T - \bar{x}_T\|\} \\
\overset{(b)}{=}& \frac{1}{2}\left(T\sqrt{\eta}\mathbb{E}\{\|\bar{y}_T - \bar{x}_T\|\} - \frac{1}{\sqrt{\eta}}\|\lambda\|\right)^2 - \frac{\|\lambda\|^2}{2\eta}
\end{aligned}$$

where (a) follows from the Cauchy-Schwarz and Jensen's inequalities, and (b) is obtained by completing the square.

$$\begin{aligned}
\mathbf{LHS}_b =& \frac{G^2\beta}{2}\sum_{t=1}^{T-1}\mathbb{E}\left\{\|y_{t+1} - y_t\|^2\right\} - G\|\mu\|_2\sum_{t=1}^{T-1}\mathbb{E}\{\|y_{t+1} - y_t\|\} \\
\overset{(c)}{\geq}& \frac{G^2\beta}{2}(T-1)\left(\frac{1}{T-1}\sum_{t=1}^{T-1}\mathbb{E}\{\|y_{t+1} - y_t\|\}\right)^2 \\
& - G\|\mu\|_2\sum_{t=1}^{T-1}\mathbb{E}\{\|y_{t+1} - y_t\|\} \\
\overset{(d)}{=}& \frac{1}{2}\left(\sqrt{\frac{G^2\beta}{T-1}}\sum_{t=1}^{T-1}\mathbb{E}\{\|y_{t+1} - y_t\|\} - \sqrt{\frac{T-1}{\beta}}\|\mu\|_2\right)^2 - \frac{T-1}{2\beta}\|\mu\|_2^2
\end{aligned}$$

here, (c) is justified by Jensen's inequality, and (d) is obtained by completing the square.

$$
\begin{aligned}
\mathbf{LHS}_c =& \frac{\beta}{2}\mathbb{E}\left\{\|W_T\|_2^2 + \|[h(y_T)]_+\|_2^2\right\} - \mathbb{E}\left\{\mu^\top h(y_T) + \mu^\top W_T\right\} \\
\overset{(e)}{\geq}& \frac{\beta}{4}\left(\mathbb{E}\left\{\|W_T\|_2 + \|[h(y_T)]_+\|_2\right\}\right)^2 - \mathbb{E}\left\{\mu^\top h(y_T) + \mu^\top W_T\right\} \\
\overset{(f)}{\geq}& \frac{\beta}{4}\left(\mathbb{E}\left\{\|W_T\|_2 + \|[h(y_T)]_+\|_2\right\}\right)^2 - \|\mu\|_2\mathbb{E}\left\{\|[h(y_T)]_+\|_2 + \|W_T\|_2\right\} \\
\overset{(g)}{=}& \left(\frac{\sqrt{\beta}}{2}\mathbb{E}\left\{\|W_T\|_2 + \|[h(y_T)]_+\|_2\right\} - \frac{\|\mu\|_2}{\sqrt{\beta}}\right)^2 - \frac{\|\mu\|_2^2}{\beta}
\end{aligned}
$$

where (e) holds by Jensen's inequality; (f) holds because of the the Cauchy-Schwarz inequality and the fact that $\mu^\top h(y_T) \leq \mu^\top [h(y_T)]_+$; and (g) is by completing the square.

By plugging these three inequalities back into Eq. (A10), we get:

$$
\begin{aligned}
&\frac{1}{2}\left(T\sqrt{\eta}\mathbb{E}\left\{\|\bar{y}_T - \bar{x}_T\|\right\} - \frac{1}{\sqrt{\eta}}\|\lambda\|\right)^2 \\
&+ \frac{1}{2}\left(\sqrt{\frac{G^2\beta}{T-1}}\sum_{t=1}^{T-1}\mathbb{E}\left\{\|y_{t+1} - y_t\|\right\} - \sqrt{\frac{T-1}{\beta}}\|\mu\|_2\right)^2 \\
&+ \left(\frac{\sqrt{\beta}}{2}\mathbb{E}\left\{\|W_T\|_2 + \|[h(y_T)]_+\|_2\right\} - \frac{\|\mu\|_2}{\sqrt{\beta}}\right)^2 \\
&\leq \underbrace{\beta\|h(x^*)\|_2^2 + G^2D^2\beta + \frac{T+1}{2\beta}\|\mu\|_2^2 + \frac{TL^2}{2\alpha} + \frac{\alpha D^2}{2} + T\eta\frac{D^2+2\delta}{2} + \frac{\|\lambda\|^2}{2\eta}}_{\Gamma_T}
\end{aligned}
\tag{A11}
$$

The equation above can be employed to individually bound each of the three left-hand-side terms with respect to the newly defined parameter Γ_T. This implies:

$$
\begin{aligned}
\frac{1}{2}\left(T\sqrt{\eta}\mathbb{E}\left\{\|\bar{y}_T - \bar{x}_T\|\right\} - \frac{1}{\sqrt{\eta}}\|\lambda\|\right)^2 &\leq \Gamma_T \\
\frac{1}{2}\left(\sqrt{\frac{G^2\beta}{T-1}}\sum_{t=1}^{T-1}\mathbb{E}\left\{\|y_{t+1} - y_t\|\right\} - \sqrt{\frac{T-1}{\beta}}\|\mu\|_2\right)^2 &\leq \Gamma_T \\
\left(\frac{\sqrt{\beta}}{2}\mathbb{E}\left\{\|W_T\|_2 + \|[h(y_T)]_+\|_2\right\} - \frac{\|\mu\|_2}{\sqrt{\beta}}\right)^2 &\leq \Gamma_T
\end{aligned}
$$

Which become

$$\begin{aligned}\mathbb{E}\left\{\|\bar{y}_T-\bar{x}_T\|\right\} \leq & \sqrt{\frac{2}{T^2\eta}\Gamma_T}+\frac{\|\lambda\|}{T\eta}\\ \sum_{t=1}^{T-1}\mathbb{E}\left\{\|y_{t+1}-y_t\|\right\} \leq & \sqrt{\frac{2T}{G^2\beta}\Gamma_T}+\frac{T\|\mu\|_2}{G\beta}\\ \mathbb{E}\left\{\|W_T\|_2+\|\left[h(y_T)\right]_+\|_2\right\} \leq & \sqrt{\frac{4}{\beta}\Gamma_T}+\frac{2\|\mu\|_2}{\beta}\end{aligned} \tag{A12}$$

Substituting (A12) in Lemma 6 yields:

$$\begin{aligned}\mathbb{E}\left\{\|\left[h(\bar{x}_T)\right]_+\|_2\right\} \leq & \frac{G}{T}\left(\sqrt{\frac{2T}{G^2\beta}\Gamma_T}+\frac{T\|\mu\|_2}{G\beta}\right)\\ & +\frac{1}{T}\left(\sqrt{\frac{4}{\beta}\Gamma_T}+\frac{2\|\mu\|_2}{\beta}\right)\\ & +G\left(\sqrt{\frac{2}{T^2\eta}\Gamma_T}+\frac{\|\lambda\|}{T\eta}\right)\\ = & \sqrt{\frac{\Gamma_T}{T}}\left(\sqrt{\frac{2}{\beta}}+\sqrt{\frac{4}{T\beta}}+\sqrt{\frac{2G^2}{T\eta}}\right)\\ & +\frac{\|\mu\|_2}{\beta}+\frac{2\|\mu\|_2}{T\beta}+\frac{G\|\lambda\|}{T\eta},\end{aligned} \tag{A13}$$

Parameter Selection 1 (ParSel.1): By substituting $\eta=\epsilon, \alpha=\beta=1/\epsilon$, and $T\geq 1/\epsilon^2$ into (A13) and utilizing the Γ_T defined in (A11), we obtain:

$$\frac{\Gamma_T}{T}\leq \mathcal{O}(\epsilon),$$

and thus

$$\mathbb{E}\left\{\|\left[h(\bar{x}_T)\right]_+\|_2\right\}\leq \mathcal{O}(\epsilon).$$

Parameter Selection 2 (ParSel.2): To proceed, we must first simplify (A13) further.

$$\begin{aligned}\mathbb{E}\left\{\|\left[h(\bar{x}_T)\right]_+\|_2\right\} \overset{(a)}{\leq} & \sqrt{\frac{\Gamma_T}{T}}\left(\left(1+\sqrt{2}\right)\sqrt{\frac{2}{\beta}}+\sqrt{\frac{2G^2}{T\eta}}\right)+\frac{3\|\mu\|_2}{\beta}+\frac{G\|\lambda\|}{T\eta}\\ = & \sqrt{\frac{\Gamma_T}{T}\left(1+\sqrt{2}\right)^2\frac{2}{\beta}}+\sqrt{\frac{\Gamma_T}{T}\frac{2G^2}{T\eta}}+\sqrt{\frac{9\|\mu\|_2^2}{\beta^2}}+\sqrt{\frac{G^2\|\lambda\|^2}{T^2\eta^2}}\\ \overset{(b)}{\leq} & \sqrt{\frac{\Gamma_T}{T}\left(\frac{47}{\beta}+\frac{8G^2}{T\eta}\right)+\frac{36\|\mu\|_2^2}{\beta^2}+\frac{4G^2\|\lambda\|^2}{T^2\eta^2}}\end{aligned}$$

$$\overset{(c)}{\leq}\sqrt{\frac{\Gamma_T}{T}\left(\frac{47}{\beta}+\frac{8G^2}{T\eta}\right)+\frac{36\|\mu\|_2^2}{\beta^2}+\frac{4G^2\left(L+G\|\mu\|_2\right)^2}{T^2\eta^2}} \tag{A14}$$

here, for (a), we exploit the fact that $T \geq 1$; for (b), we apply Jensen's inequality to the concave square root function; and finally, (c) follows from (5).

Simplifying the term $\sqrt{\frac{\Gamma_T}{T}}$ using the definition from (A11) results in:

$$\begin{aligned}\frac{\Gamma_T}{T} =& \frac{\beta\|h(x^*)\|_2^2}{T}+\frac{G^2D^2\beta}{T}+\frac{T+1}{2\beta T}\|\mu\|_2^2+\frac{L^2}{2\alpha}+\frac{\alpha D^2}{2T}+\eta\frac{D^2+2\delta}{2}+\frac{\|\lambda\|^2}{2T\eta}\\ \overset{(a)}{\leq}& \frac{\beta\|h(x^*)\|_2^2}{T}+\frac{G^2D^2\beta}{T}+\frac{\|\mu\|_2^2}{\beta}+\frac{L^2}{2\alpha}+\frac{\alpha D^2}{2T}+\eta\frac{D^2+2\delta}{2}+\frac{\|\lambda\|^2}{2T\eta}\\ \overset{(b)}{\leq}& \frac{\beta\|h(x^*)\|_2^2}{T}+\frac{G^2D^2\beta}{T}+\frac{\|\mu\|_2^2}{\beta}+\frac{L^2}{2\alpha}+\frac{\alpha D^2}{2T}+\eta\frac{D^2+2\delta}{2}+\frac{(L+G\|\mu\|_2)^2}{2T\eta}\end{aligned} \tag{A15}$$

where (a) uses the fact that $T \geq 1$, and (b) is satisfied based on the inequality (5). Replacing $\alpha = \frac{L\sqrt{T}}{D}$, $\eta = \frac{L}{\sqrt{T(D^2+2\delta)}}$, and $\beta = \frac{\sqrt{T}}{GD}$ in (A15) and (A14) we get:

$$\begin{aligned}\frac{\Gamma_T}{T} \leq& \frac{1}{\sqrt{T}}\Bigg(LD+L\sqrt{D^2+2\delta}+GD+\frac{\|h(x^*)\|_2^2}{GD}\\ &+\|\mu\|_2G\sqrt{D^2+2\delta}+\|\mu\|_2^2\left(GD+\frac{G^2}{2L}\sqrt{D^2+2\delta}\right)\Bigg).\end{aligned}$$

and thus

$$\mathbb{E}\left\{\|\left[h(\bar{x}_T)\right]_+\|_2\right\} \leq \frac{1}{\sqrt{T}}\sqrt{A_0+A_1\|\mu\|_2+A_2\|\mu\|_2^2}$$

where A_0, A_1 and A_2 are defined as follows:

$$\begin{aligned}A_2:=&55\frac{G^3}{L}D\sqrt{D^2+2\delta}+83G^2D^2+8\frac{G^4}{L^2}(D^2+2\delta)\\ A_1:=&16\frac{G^3}{L}(D^2+2\delta)+47G^2D\sqrt{D^2+2\delta}\\ A_0:=&47GLD^2+47GLD\sqrt{D^2+2\delta}+47G^2D^2+12G^2(D^2+2\delta)\\ &+8G^2D\sqrt{D^2+2\delta}+8\frac{G^3}{L}D\sqrt{D^2+2\delta}+\left\|h(x^*)\right\|_2^2\left(47+8\frac{G}{L}\frac{\sqrt{D^2+2\delta}}{D}\right)\end{aligned}$$

□

Remark 4 If the functional constraints satisfy the assumption that for all $i \in \{1, \ldots, m\}$ there exists a point $z_i \in \mathcal{X}$ such that $h_i(z_i) \geq 0$ (meaning that none of the functional inequalities are strictly satisfied everywhere on the set $\mathcal{X}$), then we can write:

$$\|h(x^*)\|_2 \leq GD.$$

Proof Using Lemma 1 we have:

$$h_i(z) - h_i(x^*) \leq G_i \|z_i - x^*\|.$$

Assumption 1 implies $\|z_i - x^*\| \leq D$, thus

$$h_i(z) - h_i(x^*) \leq G_i D.$$

Using the fact that $h_i(x^*) \leq 0$ and $h_i(z_i) \geq 0$, we get

$$(h(x^*))^2 \leq G_i^2 D^2.$$

Finally, summing over $i = \{1, \ldots, m\}$ gives

$$\left\|h(x^*)\right\|_2^2 \leq \sum_{i=1}^{m} G_i^2 D^2 \leq G^2 D^2$$

where we used Assumption 3.iii in the last step. □

A.2 Upper bound for $\|Q_t\|$

Here, we prove an upper bound for $\|Q_t\|$ as a secondary result of Eq. (A12). Although this bound does not explicitly appear in our guarantees, it is useful when using an inexact LMO. For a fixed parameter δ, the computational complexity of Line 6 of the algorithm, IN- LMO$_{\mathcal{X}}\{-Q_t; \delta\}$, depends on the size of Q_t.

Theorem 3 *(Upper Bound for $\|Q_t\|$) Given Assumptions 1 to 4, for Algorithm 1, with any $T \in \{1, 2, 3, \ldots\}$, $\eta > 0$, $\alpha > 0$, $\beta > 0$, and $\delta \geq 0$, the expected value of $\|Q_t\|$ for any $t \in \{1, 2, 3, \ldots\}$ is bounded as follows:*

$$\begin{aligned}\mathbb{E}\{\|Q_t\|\} \leq & \sqrt{t}\sqrt{\frac{2\|\mu\|_2^2}{\beta\eta} + \frac{L^2}{\alpha\eta} + D^2 + 2\delta} \\ & + \sqrt{\frac{2\beta}{\eta}\|h(x^*)\|_2^2 + \frac{2G^2D^2\beta}{\eta} + \frac{\alpha D^2}{\eta} + \frac{(L + G\|\mu\|_2)^2}{\eta^2}} + \frac{\|\lambda\|}{\eta}.\end{aligned} \tag{A16}$$

In particular, under Parameter Selection (ParSel.1) we have

$$\mathbb{E}\{\|Q_t\|\} \leq \mathcal{O}\left(\sqrt{t} + \frac{1}{\epsilon}\right), \tag{A17}$$

while under Parameter Selection (ParSel.2) we have

$$\mathbb{E}\{\|Q_t\|\} \leq B_1\sqrt{t} + B_2\sqrt{T}. \tag{A18}$$

Here, B_1 and B_2 are constants that depend on the problem's parameters, and they are defined in the final part of the theorem's proof.

Proof Consider Eqs. (A12) and (A15). Throughout their derivation, no assumptions were made about T, meaning it can be replaced with any positive integer, regardless of the number of iterations fixed in the algorithm. Thus, for all $t \in \{1, 2, \ldots\}$, the first part of (A12) gives:

$$\frac{1}{t}\mathbb{E}\{\|Q_t\|\} = \mathbb{E}\{\|\bar{y}_t - \bar{x}_t\|\} \leq \sqrt{\frac{2}{t^2\eta}\Gamma_t} + \frac{\|\lambda\|}{t\eta}$$

$$\implies \quad \mathbb{E}\{\|Q_t\|\} \leq \sqrt{\frac{2\Gamma_t}{\eta}} + \frac{\|\lambda\|}{\eta}$$

Replacing T in (A15) with t, we have:

$$\Gamma_t \leq \beta\|h(x^*)\|_2^2 + G^2D^2\beta + t\frac{\|\mu\|_2^2}{\beta} + t\frac{L^2}{2\alpha} + \frac{\alpha D^2}{2} + \eta t\frac{D^2+2\delta}{2} + \frac{(L+G\|\mu\|_2)^2}{2\eta}$$

Combining the above equations and using Jensen's inequality gives (A16).

Parameter Selection 1 (ParSel.1): By substituting $\eta = \epsilon, \alpha = \beta = 1/\epsilon$ into (A16), we obtain (A17).

Parameter Selection 2 (ParSel.2): Plugging the values $\alpha = \frac{L\sqrt{T}}{D}, \eta = \frac{L}{\sqrt{T(D^2+2\delta)}}$, and $\beta = \frac{\sqrt{T}}{GD}$ in (A16) gives (A18), where B_1 and B_2 are defined as follows:

$$B_1 := \sqrt{\frac{D}{L}\left(L + 2G\|\mu\|_2^2\right)\sqrt{D^2+2\delta} + D^2 + 2\delta}$$

$$B_2 := \frac{\|\lambda\|}{L}\sqrt{D^2+2\delta} + \sqrt{\sqrt{D^2+2\delta}\left(\frac{2}{GLD}\|h(x^*)\|_2^2 + \frac{2GD}{L} + D\right) + \frac{(L+G\|\mu\|_2)^2\left(D^2+2\delta\right)}{L^2}}$$

□

Appendix B Nuclear norm

The bounded nuclear norm domain is a well-known example where inexact linear minimization holds a significant computational advantage over projection [12, 68]. In this section, we provide an overview of some key results regarding this norm and its characteristics.

Definition 2 Singular Value Decomposition (SVD) of a real valued $q \times p$ matrix c is a factorization of the form $c = u\sigma v^\top$, where u is an $q \times q$ orthogonal matrix, σ is an diagonal matrix with nonnegative real numbers on the diagonal, and v is an $p \times p$ orthogonal matrix. The diagonal entries $\sigma_i := \sigma_{i,i}$ are uniquely determined by c and are known as the singular values of c. Let the i-th column of matrix u be denoted as u_i, and the i-th column of matrix v be denoted as v_i.

Definition 3 Nuclear norm or trace norm of a $q \times p$ matrix c with SVD of form $c = u\sigma v^\top$, is defined as:

$$\|c\|_* = \sum_{i=1}^{\min\{q,p\}} \sigma_i$$

Similar to the definition in Problem (R4NR), for a fixed $\gamma > 0$, let the set $\mathcal{C}_\gamma$ denote all real-valued $q \times p$ matrices with a nuclear norm smaller than γ.

Lemma 7 *The nuclear norm of a matrix can be lower bounded by its norm:*

$$\|c\|_* \geq \|c\| := \mathrm{Tr}(c^\top c)$$

Proof See [103]. □

This lemma provides an upper bound on the Euclidean diameter of the set $\mathcal{C}_\gamma$ (denoted as the parameter D in the algorithm described in Sect. 4.1).

$$\|c_t - c^*\| \leq \|c_t\| + \|c^*\| \leq \|c_t\|_* + \|c^*\|_* \leq 2\gamma \tag{19}$$

We implemented the exact and inexact LMOs used in Sect. 4.1 based on the following lemmas.

Lemma 8 *(Linear optimization on nuclear norm-ball) Let z be a $q \times p$ matrix and with singular-value decomposition $z = u\sigma v^\top$. Then*

$$\gamma \cdot u_1 v_1^\top \in \arg\min_{c \in \mathcal{C}_\gamma} \langle c\,,\, z\rangle\,.$$

Here, the u_1, v_1 are the vectors corresponding to the largest singular value.

Proof See [12]. □

The projection oracle employed in the PGD algorithm in Sect. 4.1 utilizes the following lemma.

Lemma 9 *(Projection onto the nuclear norm-ball) Let z be a $q \times p$ matrix and consider its singular-value decomposition $a = u\sigma v^\top$. If $\|z\|_* \geq \gamma$, then the Euclidean projection of z onto the nuclear norm-ball with radius γ is given by*

$$\mathrm{PO}_{\mathcal{C}_\gamma}\{c\} = \sum_{i=1}^{\min\{q,p\}} \max\{0, \sigma_i - \zeta\} u_i v_i^\top,$$

where $\zeta \geq 0$ is the solution to equation

$$\sum_{i=1}^{\min\{q,p\}} \max\{0, \sigma_i - \zeta\} = \gamma.$$

Proof See [6, 104]. □

Appendix C Further insights into inexact LMOs

The study in [9], particularly Table 1, provides a comprehensive comparison of the computational complexities between inexact linear optimization and projections onto various significant sets.

Nuclear norm-ball: Implementing the oracle IN- $\mathrm{LMO}_{\mathcal{C}_\gamma}\{-Q; \delta\}$ for a nuclear norm-ball (the set $\mathcal{C}_\gamma$ defined in Problem (R4NR)) includes calculating the largest singular value of the matrix w and the corresponding vectors (see Lemma 8). This computation can be performed using the Lanczos algorithm [105, 106]. The computational cost of running IN- $\mathrm{LMO}_{\mathcal{C}_\gamma}\{-Q; \delta\}$ with this algorithm is

$$\mathcal{O}\left(\sqrt{\gamma}\,\Phi(Q)\log(q+p)\frac{\sqrt{\sigma_1(Q)}}{\sqrt{\delta}}\right) \tag{C20}$$

arithmetic operations [12, 106]. Here, $\sigma_1(Q)$ denotes the largest singular value (spectral norm) of the matrix w, and $\Phi(Q)$ represents the count of nonzero elements in the matrix w.

C.1 Empirical study of the inexact LMO used in Sect. 4.1

We used `randomized_svd` from the `Scikit-learn` library to implement IN- $\mathrm{LMO}_{\mathcal{C}_\gamma}$ in Sect. 4.1. This method does not directly allow for controlling the error parameter δ, as required by Assumption 3.i. To manage the error, the parameters of this method were configured as follows: `n_oversamples=1` and `n_iter=2`, both set significantly below their default values to enhance computation speed. The resultant errors minimally impacted our algorithm, as demonstrated in the experimental results (see Figs. 1 and 2).

To evaluate whether our theoretical results can describe the strong performance observed in experiments, the following measurements were conducted. Denote the

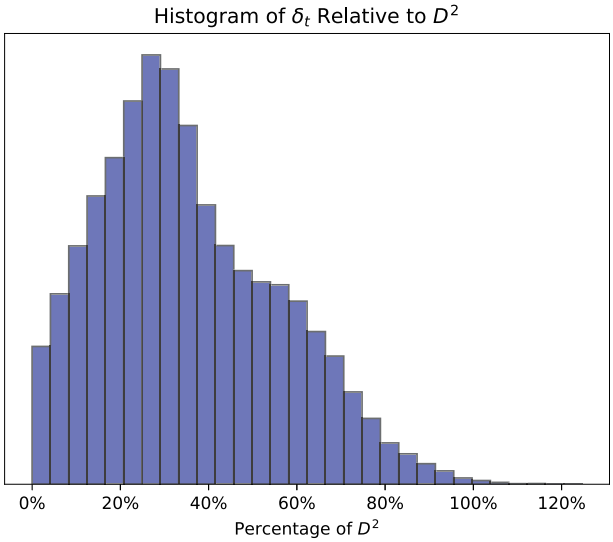

Fig. 7 Histogram of δ_t as a percentage of D^2, confirming compliance with the requirements of Remark 2 and explaining the sustained performance levels in empirical tests

coefficient matrix at iterations $t = 1, 2, \ldots, T$ as $c_{t+1} \leftarrow \text{IN-LMO}_{\mathcal{C}_\gamma}\{-Q_t; \delta\}$. Define the empirical error of the inexact linear minimization oracle $\hat{\delta}_t$ as:

$$\hat{\delta}_t := \langle c_{t+1}, -Q_t \rangle - \min_{c \in \mathcal{C}_\gamma} \langle c, -Q_t \rangle$$

Remark 2 discussed that if δ is of the order of D^2, then the performance should not see a significant drop when using an inexact LMO instead of an exact one. The parameter D in the experiments conducted here (and also in Sect. 4.1) is set to be 2γ, as established in Eq. (19). We fixed $\gamma = 350$ and $T = 4000$. The other parameters remain the same as in Sect. 4.1.

Figure 7 demonstrates the histogram of δ_t as a percentage of D^2. Clearly, these values satisfy our requirement of Remark 2 and show why no significant drop in performance is observed in experiments. It is also interesting to see if Eq. (C20) is informative. Figure 8 portrays the histogram of the empirical values $\sigma_1(Q_t)/\delta_t$. Since we fixed the parameters corresponding to the computational complexity of `randomized_svd`, we expect to see that the value δ_t is correlated to $\Phi(Q_t) \cdot \sigma_1(Q_t)$. Our observations indicate that all of the matrices Q_t are dense. Thus, ignoring the term $\Phi(Q_t)$, Fig. 8 is in line with our expectations.

The next step is to analyze the growth of $\sigma_1(Q_t)$ over time. Since the largest singular value $\sigma_1(Q_t)$ can be bounded by the norm [103]: $\sigma_1(Q_t) \leq \|Q_t\|$, Theorem 3 suggests asymptotic growth proportional to $\sqrt{t}$. Figure 9 shows a plot of $(\sigma_1(Q_t))^2$, experimentally demonstrating how this value changes over time. Initially, $\sigma_1(Q_t)$ grows at a rate of $\mathcal{O}(\sqrt{t})$, but it eventually reaches a phase where the growth stops, and the value stabilizes.

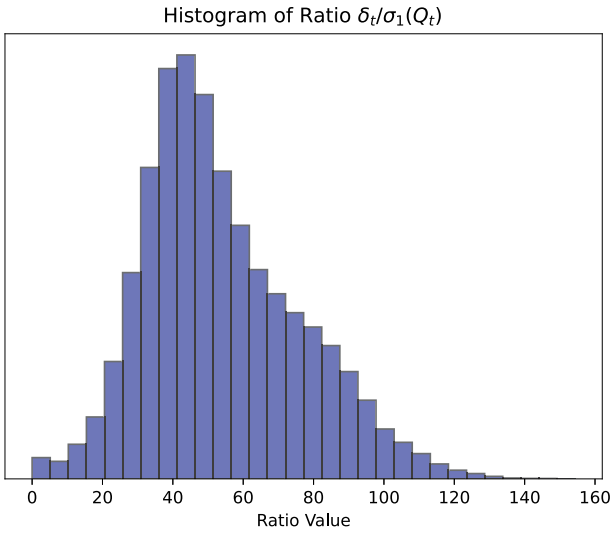

Fig. 8 Histogram of the ratio $\frac{\sigma_1(Q_t)}{\delta_t}$ under fixed computational complexity parameters, illustrating the correlation between δ_t and the largest singular value in matrices Q_t

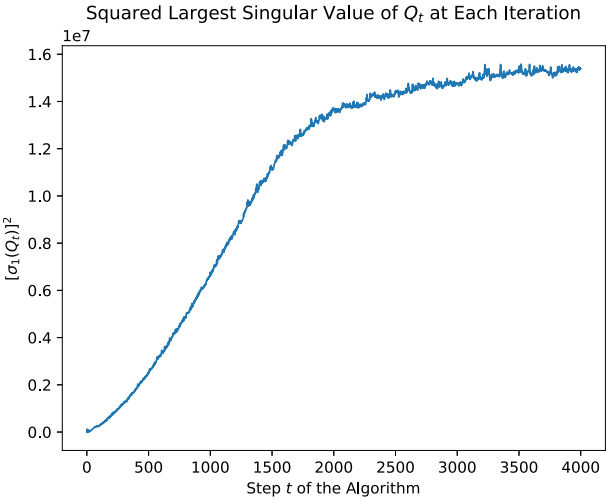

Fig. 9 Plot of $(\sigma_1(Q_t))^2$ over time, showing the initial growth of $\sigma_1(Q_t)$ as $\mathcal{O}(\sqrt{t})$ before convergence, indicating a phase shift in its behavior

Appendix D Remarks on choosing the auxiliary set

An important assumption made in this paper is that, on a chosen set $\mathcal{Y}$, both the objective and constraint functions are convex, Lipschitz continuous, and that we have access to their (stochastic) subgradients (see Lemma 1 and Assumption 3.iii). In some problems, such as the one considered in Sect. 4.1 and 4.2, these assumptions are

satisfied over the entire space $\mathcal{Y} = \mathbb{V}$. However, there can be instances where this assumption does not hold.

One way to address this issue is to select a set $\mathcal{Y} \neq \mathbb{V}$ that still meets the requirements of Assumption 3.ii. Considering that Assumption 1 implies $\mathcal{X}$ is bounded, the two simplest choices for $\mathcal{Y}$ are a sphere centered at x_1 with a radius of at most D, and a hypercube centered at x_1 with an edge length of at most D.

While this method covers many examples, there remain some cases where things can go wrong. For instance, the Lipschitz constant over the set $\mathcal{Y}$ might be significantly larger than its value over the set $\mathcal{X}$, which can result in a drop in algorithm performance. Additionally, we may simply not have access to the functions outside of the set $\mathcal{X}$. As mentioned in Sect. 5, it remains an open question whether there exists an algorithm for this setup that only requires the subgradient of points belonging to $\mathcal{X}$.

A possible solution to this problem is to extend the functions from $\mathcal{X}$ to a set $\mathcal{Y}$ (such as $\mathcal{Y} = \mathbb{V}$), while maintaining convexity and Lipschitz continuity without increasing the Lipschitz constant. The existence of such extension is provided by the *McShane-Whitney extension theorem* [82].[11] This extension was used in the proof of Lemma 2, but here it can also serve as an analytical preprocessing step. Consider f to be a convex and L-Lipschitz continuous function on $\mathcal{X}$. One general construction of an extension of f is by solving the following equation for every $x \in \mathbb{V}$ [82, 109]:

$$\tilde{f}(x) := \inf_{z \in \mathcal{X}} \{f(z) + L\|x - z\|\}. \tag{D21}$$

Here, the new function $\tilde{f}$ extends f from $\mathcal{X}$ to the entire $\mathbb{V}$, which means it satisfies the followings:

- $\tilde{f}$ is convex and L-Lipschitz continuous on $\mathbb{V}$.
- $\tilde{f}(x) = f(x), \forall x \in \mathcal{X}$.

Notice that solving an optimization problem as described in Eq. (D21) is at least as challenging as performing a projection onto the set $\mathcal{X}$. This suggests that part of our algorithm's effectiveness may arise from leveraging additional information that is often available: the *subgradient of the function outside of the feasible set*. The Projected Gradient Descent algorithm does not utilize this extra information; however, as mentioned in Sect. 5, our algorithm is not unique in doing so.

An Example This simple example provides intuition on what was discussed in this section. Consider the following one-dimensional problem:

$$\begin{aligned} \text{Minimize:} \quad & f(x) := \max\{\exp(-x), \exp(x)\} \\ \text{Subject to:} \quad & x \in \mathcal{X} := \{x \in \mathbb{R} : -1 \leq x \leq 1\} \end{aligned}$$

The function f is convex and e-Lipschitz continuous on $\mathcal{X}$. While this function remains convex on $\mathbb{R}$, it is not Lipschitz continuous. Suppose an initial feasible point $x_1 = \frac{1}{2}$ is given. The diameter of $\mathcal{X}$ is 2. Choosing $\mathcal{Y}$ as prescribed in this section results in $\mathcal{Y} =$

[11] We only use part of the McShane-Whitney extension theorem that deals with extending convex functions. This theorem can also extend other kinds of Lipschitz continuous functions. For more details on extending convex functions, see [107, 108].

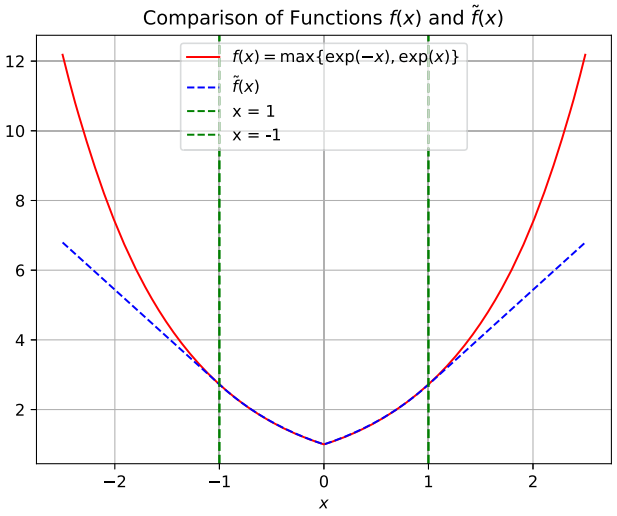

Fig. 10 Comparison of the original function $f(x)$, defined as $\max\{\exp(-x), \exp(x)\}$, with its extended version $\tilde{f}(x)$, when restricted to $x \in [-1, 1]$. This figure illustrates how $\tilde{f}$ maintains the e-Lipschitz continuity beyond the original domain $\mathcal{X}$, contrasting with $f(x)$ which is not Lipschitz continuous on $\mathbb{R}$

$\{x \in \mathbb{R} : -2 \leq |x - x_1| \leq 2\}$. The function f becomes $e^{5/2}$-Lipschitz continuous on $\mathcal{Y}$. To avoid this increase in the Lipschitz constant, we can derive an extended function $\tilde{f}$ using Eq. (D21):

$$\tilde{f}(x) = \inf_{z \in \mathcal{X}} \{f(z) + e|x - z|\} = \begin{cases} e \cdot x & \text{if } 1 < x \\ f(x) & \text{if } -1 \leq x \leq 1 \\ -e \cdot x & \text{if } x < -1 \end{cases}$$

Figure 10 shows these two functions.

References

1. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004). https://doi.org/10.1017/CBO9780511804441
2. Palomar, D.P., Eldar, Y.C.: Convex Optimization in Signal Processing and Communications. Cambridge University Press, Cambridge (2009). https://doi.org/10.1017/CBO9780511804458
3. Sra, S., Nowozin, S., Wright, S.J.: Optimization for Machine Learning. The MIT Press, Cambridge (2012)
4. Nesterov, Y., Nemirovskii, A.: Interior-Point Polynomial Algorithms in Convex Programming. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1994). https://doi.org/10.1137/1.9781611970791
5. Nesterov, Y.: Lectures on Convex Optimization, 2nd ed. 2018. edn. Springer Optimization and Its Applications, 137. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91578-4
6. Beck, A.: First-order Methods in Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2017). https://doi.org/10.1137/1.9781611974997
7. Lan, G.: First-order and Stochastic Optimization Methods for Machine Learning vol. 1. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39568-1

8. Perez, G., Ament, S., Gomes, C., Barlaud, M.: Efficient projection algorithms onto the weighted $l1$ ball. Artif. Intell. **306**, 103683 (2022)
9. Combettes, C.W., Pokutta, S.: Complexity of linear minimization and projection on some sets. Oper. Res. Lett. **49**(4), 565–571 (2021). https://doi.org/10.1016/j.orl.2021.06.005
10. Combettes, C.W.: Frank-Wolfe Methods for Optimization and Machine Learning. PhD thesis, Georgia Institute of Technology (2021)
11. Juditsky, A., Nemirovski, A.: Solving variational inequalities with monotone operators on domains given by linear minimization oracles. Math. Program. **156**(1–2), 221–256 (2016)
12. Jaggi, M.: Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 28, pp. 427–435. PMLR, Atlanta, Georgia, USA (2013)
13. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval Res. Logist Q. **3**(1–2), 95–110 (1956). https://doi.org/10.1002/nav.3800030109
14. Argyriou, A., Signoretto, M., Suykens, J.: Hybrid Conditional Gradient–Smoothing Algorithms with Applications to Sparse and Low Rank Regularization (2014)
15. Levitin, E.S., Polyak, B.T.: Constrained minimization methods. USSR Comput. Math. Math. Phys. **6**(5), 1–50 (1966). https://doi.org/10.1016/0041-5553(66)90114-5
16. Yurtsever, A., Fercoq, O., Locatello, F., Cevher, V.: A Conditional Gradient Framework for Composite Convex Minimization with Applications to Semidefinite Programming (2018)
17. Lan, G.: The Complexity of Large-scale Convex Programming under a Linear Optimization Oracle (2014)
18. Nemirovskij, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization (1983)
19. Bubeck, S., *et al.*: Convex optimization: algorithms and complexity. Foundations and Trends® in Machine Learning 8(3-4), 231–357 (2015)
20. Nemirovski, A.: Information-based complexity of convex programming. Lecture Notes **834** (1995)
21. Lacoste-Julien, S., Jaggi, M., Schmidt, M., Pletscher, P.: Block-coordinate Frank-Wolfe optimization for structural SVMs. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 28, pp. 53–61. PMLR, Atlanta, Georgia, USA (2013)
22. Jing, N., Fang, E.X., Tang, C.Y.: Robust matrix estimations meet frank–wolfe algorithm. Machine Learning, 1–38 (2023)
23. Mu, C., Zhang, Y., Wright, J., Goldfarb, D.: Scalable robust matrix recovery: frank-wolfe meets proximal methods. SIAM J. Sci. Comput. **38**(5), 3291–3317 (2016)
24. Combettes, C.W., Pokutta, S.: Revisiting the approximate carathéodory problem via the frank-wolfe algorithm. Math. Program. **197**(1), 191–214 (2023)
25. Hazan, E., Kakade, S.M., Singh, K., Soest, A.V.: Provably Efficient Maximum Entropy Exploration (2019)
26. Lin, J.-L., Hung, W., Yang, S.-H., Hsieh, P.-C., Liu, X.: Escaping from Zero Gradient: Revisiting Action-Constrained Reinforcement Learning via Frank-Wolfe Policy Optimization (2021)
27. Garber, D.: Faster Projection-free Convex Optimization over the Spectrahedron (2016)
28. Nesterov, Y.: Complexity bounds for primal-dual methods minimizing the model of objective function. Math. Program. **171**(1), 311–330 (2018). https://doi.org/10.1007/s10107-017-1188-6
29. White, D.J.: Extension of the frank-wolfe algorithm to concave nondifferentiable objective functions. J. Optim. Theory Appl. **78**(2), 283–301 (1993). https://doi.org/10.1007/BF00939671
30. Ravi, S.N., Collins, M.D., Singh, V.: A Deterministic Nonsmooth Frank Wolfe Algorithm with Coreset Guarantees (2017)
31. Cheung, E., Li, Y.: Solving separable nonsmooth problems using frank-wolfe with uniform affine approximations. In: IJCAI, pp. 2035–2041 (2018)
32. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bulletin de la Société Mathématique de France 93, 273–299 (1965) https://doi.org/10.24033/bsmf.1625
33. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005). https://doi.org/10.1007/s10107-004-0552-5
34. Parikh, N., Boyd, S.: Proximal algorithms. Foundations and Trends®in Optimization 1(3), 127–239 (2014) https://doi.org/10.1561/2400000003
35. Yurtsever, A., Tran Dinh, Q., Cevher, V.: A universal primal-dual convex optimization framework. Adv. Neural Inform. Process. Syst.**28** (2015)

36. Duchi, J.C., Bartlett, P.L., Wainwright, M.J.: Randomized smoothing for stochastic optimization. SIAM J. Optim. **22**(2), 674–701 (2012). https://doi.org/10.1137/110831659
37. Hazan, E., Kale, S.: Projection-free online learning. In: Proceedings of the 29th International Coference on International Conference on Machine Learning. ICML'12, pp. 1843–1850. Omnipress, Madison, WI, USA (2012)
38. Thekumparampil, K.K., Jain, P., Netrapalli, P., Oh, S.: Projection efficient subgradient method and optimal nonsmooth frank-wolfe method. Adv. Neural. Inf. Process. Syst. **33**, 12211–12224 (2020)
39. Polyak, V., Tret'yakov, N.: The method of penalty estimates for conditional extremum problems. USSR Comput. Math. Math. Phys. **13**(1), 42–58 (1973)
40. Lan, G., Zhou, Z.: Algorithms for stochastic optimization with function or expectation constraints. Comput. Optim. Appl. **76**(2), 461–498 (2020). https://doi.org/10.1007/s10589-020-00179-x
41. Lin, Q., Ma, R., Yang, T.: Level-set methods for finite-sum constrained convex optimization. In: International Conference on Machine Learning, pp. 3112–3121 (2018). PMLR
42. Lin, Q., Nadarajah, S., Soheili, N.: A level-set method for convex optimization with a feasible solution path. SIAM J. Optim. **28**(4), 3290–3311 (2018)
43. Hamedani, E.Y., Aybat, N.S.: A primal-dual algorithm with line search for general convex-concave saddle point problems. SIAM J. Optim. **31**(2), 1299–1329 (2021)
44. Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena Scientific, Raleigh (1999)
45. Xu, Y.: Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. Math. Program. **185**, 199–244 (2021)
46. Neely, M.J., Yu, H.: Lagrangian Methods for $\mathcal{O}(1/t)$ Convergence in Constrained convex programs. Theory, Methods, and Applications, edited by Arto Ruud, Nova Publishers, Convex Optimization (2019)
47. Lan, G., Monteiro, R.D.: Iteration-complexity of first-order augmented lagrangian methods for convex programming. Math. Program. **155**(1–2), 511–547 (2016)
48. Wei, X., Yu, H., Ling, Q., Neely, M.J.: Solving Non-smooth Constrained Programs with Lower Complexity than $\mathcal{O}(1/\epsilon)$: a Primal-Dual Homotopy Smoothing Approach (2018)
49. Yu, H., Neely, M.J.: A primal-dual type algorithm with the $\mathcal{O}(1/t)$ convergence rate for large scale constrained convex programs. In: 2016 IEEE 55th Conference on Decision and Control (CDC), pp. 1900–1905 (2016). https://doi.org/10.1109/CDC.2016.7798542
50. Yu, H., Neely, M., Wei, X.: Online convex optimization with stochastic constraints. Adv. Neural Inform. Process. Syst. **30** (2017)
51. Lemaréchal, C., Nemirovskii, A., Nesterov, Y.: New variants of bundle methods. Math. Program. **69**, 111–147 (1995)
52. Karas, E., Ribeiro, A., Sagastizábal, C., Solodov, M.: A bundle-filter method for nonsmooth convex constrained optimization. Math. Program. **116**(1), 297–320 (2009)
53. Boob, D., Deng, Q., Lan, G.: Stochastic first-order methods for convex and nonconvex functional constrained optimization. Math. Program. **197**(1), 215–279 (2023)
54. Wei, X., Neely, M.J.: Primal-Dual Frank-Wolfe for Constrained Stochastic Programs with Convex and Non-convex Objectives (2018)
55. Lan, G., Romeijn, E., Zhou, Z.: Conditional gradient methods for convex optimization with general affine and nonlinear constraints. SIAM J. Optim. **31**(3), 2307–2339 (2021)
56. Lee, D., Ho-Nguyen, N., Lee, D.: Projection-Free Online Convex Optimization with Stochastic Constraints (2023)
57. Mahdavi, M., Yang, T., Jin, R., Zhu, S., Yi, J.: Stochastic gradient descent with only one projection. Adv. Neural Inform. Process. Syst. **25** (2012)
58. Levy, K., Krause, A.: Projection free online learning over smooth sets. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 1458–1466 (2019). PMLR
59. Lee, Y.T., Sidford, A., Vempala, S.S.: Efficient convex optimization with membership oracles. In: Conference On Learning Theory, pp. 1292–1294 (2018). PMLR
60. Mhammedi, Z.: Efficient projection-free online convex optimization with membership oracle. In: Conference on Learning Theory, pp. 5314–5390 (2022). PMLR
61. Garber, D., Kretzu, B.: New projection-free algorithms for online convex optimization with adaptive regret guarantees. In: Conference on Learning Theory, pp. 2326–2359 (2022). PMLR
62. Lu, Z., Brukhim, N., Gradu, P., Hazan, E.: Projection-free adaptive regret with membership oracles. In: International Conference on Algorithmic Learning Theory, pp. 1055–1073 (2023). PMLR

63. Garber, D., Kretzu, B.: New Projection-free Algorithms for Online Convex Optimization with Adaptive Regret Guarantees (2023)
64. Gatmiry, K., Mhammedi, Z.: Projection-Free Online Convex Optimization via Efficient Newton Iterations (2023)
65. Garber, D., Kretzu, B.: Projection-free online exp-concave optimization. In: The Thirty Sixth Annual Conference on Learning Theory, pp. 1259–1284 (2023). PMLR
66. Grimmer, B.: Radial Duality Part II: Applications and Algorithms (2022)
67. Grimmer, B.: Radial Duality Part I: Foundations (2023)
68. Combettes, C.W., Pokutta, S.: Complexity of Linear Minimization and Projection on Some Sets (2021)
69. Bertsekas, D.: Convex Optimization Theory, vol. 1. Athena Scientific, Raleigh (2009)
70. Bertsekas, D.: Convex Optimization Algorithms. Athena Scientific, Raleigh (2015)
71. Kelly, F.P., Maulloo, A.K., Tan, D.K.H.: Rate control for communication networks: shadow prices, proportional fairness and stability. J. Oper. Res. Soc. **49**(3), 237–252 (1998)
72. Bertsekas, D.P.: Linear Network Optimization: Algorithms and Codes. MIT press, Cambridge (1991)
73. LeBlanc, L.J., Helgason, R.V., Boyce, D.E.: Improved efficiency of the frank-wolfe algorithm for convex network programs. Transp. Sci. **19**(4), 445–462 (1985)
74. Bertsekas, D.P.: Netw. Optim. Contin. Discrete Models. Athena Scientific, Raleigh (1998)
75. Neely, M.J.: Stochastic network optimization with application to communication and queueing systems. Synth. Lect. Commun. Netw. **3**(1), 1–211 (2010)
76. Hazan, E.: Sparse approximate solutions to semidefinite programs. In: Latin American Symposium on Theoretical Informatics, pp. 306–316 (2008). Springer
77. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust statistics. Wiley Series in Probability and Statistics (2005)
78. Huber, P.J.: In: Kotz, S., Johnson, N.L. (eds.) Robust Estimation of a Location Parameter, pp. 492–518. Springer, New York, NY (1992). https://doi.org/10.1007/978-1-4612-4380-9_35
79. Lerasle, M.: Selected topics on robust statistical learning theory. Lecture Notes (2019)
80. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
81. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: the Lasso and Generalizations. CRC press, Boca Raton, Florida (2015).https://doi.org/10.1201/b18401
82. Petrakis, I.: McShane-Whitney extensions in constructive analysis. Logical Methods Comput. Sci. (2020). https://doi.org/10.23635/LMCS-16(1:18)2020
83. Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using bregman functions. SIAM J. Optim. **3**(3), 538–543 (1993)
84. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
85. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. Submitt. SIAM J. Optim. **2**(3) (2008)
86. Borchani, H., Varando, G., Bielza, C., Larranaga, P.: A survey on multi-output regression. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **5**(5), 216–233 (2015)
87. Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE Trans. Knowl. Data Eng. **34**(12), 5586–5609 (2021)
88. Anderson, T.W.: Estimating linear restrictions on regression coefficients for multivariate normal distributions. The Annals of Mathematical Statistics, 327–351 (1951)
89. Fazel, M.: Matrix Rank Minimization with Applications. PhD thesis, PhD thesis, Stanford University (2002)
90. Chen, K., Dong, H., Chan, K.-S.: Reduced rank regression via adaptive nuclear norm penalization. Biometrika **100**(4), 901–920 (2013)
91. Ding, C., Zhou, D., He, X., Zha, H.: R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 281–288 (2006)
92. Ming, D., Ding, C., Nie, F.: A probabilistic derivation of lasso and l12-norm feature selections. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 4586–4593 (2019)
93. Thekumparampil, K.K.: Optimal nonsmooth frank-wolfe method for stochastic regret minimization. (2020). https://api.semanticscholar.org/CorpusID:229347497
94. Fratta, L., Gerla, M., Kleinrock, L.: The flow deviation method: an approach to store-and-forward communication network design. Networks **3**(2), 97–133 (1973)

95. Klessig, R.W.: An algorithm for nonlinear multicommodity flow problems. Networks **4**(4), 343–355 (1974)
96. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. MIT press, Cambridge, MA (2022). https://doi.org/10.5555/1614191
97. Kovács, P.: Minimum-cost flow algorithms: an experimental evaluation. Optim. Methods Softw. **30**(1), 94–127 (2015)
98. Blair, C.: Problem complexity and method efficiency in optimization (a. s. nemirovsky and d. b. yudin). SIAM Review 27(2), 264–265 (1985) https://doi.org/10.1137/1027074
99. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. **31**(3), 167–175 (2003)
100. Bach, F., Moulines, E.: Non-strongly-convex smooth stochastic approximation with convergence rate $\mathcal{O}(1/n)$ (2013)
101. Hazan, E., Minasyan, E.: Faster projection-free online learning. In: Conference on Learning Theory, pp. 1877–1893 (2020). PMLR
102. Tao, W., Pan, Z., Wu, G., Tao, Q.: The strength of nesterov's extrapolation in the individual convergence of nonsmooth optimization. IEEE Transactions on Neural Networks and Learning Systems, 1–12 (2019) https://doi.org/10.1109/tnnls.2019.2933452
103. Golub, G.H., Van Loan, C.F.: Matrix Computations. JHU press, Baltimore (2013). https://doi.org/10.56021/9781421407944
104. Garber, D.: On the convergence of projected-gradient methods with low-rank projections for smooth convex minimization over trace-norm balls and related problems. SIAM J. Optim. **31**(1), 727–753 (2021)
105. Lanczos, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators (1950)
106. Kuczyński, J., Woźniakowski, H.: Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. SIAM J. Matrix Anal. Appl. **13**(4), 1094–1122 (1992)
107. Dragomirescu, F., Ivan, C.: The smallest convex extensions of a convex function. Optimization **24**(3–4), 193–206 (1992)
108. Yan, M.: Extension of convex function. J. Convex Anal. **21**(4), 965 (2014)
109. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis, vol. 317. Springer, Heidelberg, Berlin, New York (2009)