Nonsmooth Projection-Free Optimization with Functional Constraints

Kamiar Asgari and Michael J. Neely

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

*Corresponding author(s). E-mail(s): Kamiaras@usc.edu;
Contributing authors: Mjneely@usc.edu;

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 $\boldsymbol{\epsilon}$-suboptimal solution in $\boldsymbol{\mathcal{O}(\epsilon^{-2})}$ iterations, with each iteration requiring only a single (potentially inexact) Linear Minimization Oracle (LMO) 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

MSC Classification: 65K05 , 65K10 , 65K99 , 90C25 , 90C15 , 90C25 , 90C30.

1 Introduction

Set $\mathbb{V}$ to 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 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 (possibly nonsmooth) 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 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 demands $\mathcal{O}(\epsilon^{-2})$ queries to a LMO, which was proven to be optimal [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 a 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].

With respect to functional inequality constraints, prior work explores various techniques, including cooperative subgradients [39], level-set [40, 41], exact penalty and augmented Lagrangian methods [42–45], and Lyapunov drift-plus-penalty [46–49].

The Frank-Wolfe algorithm has been generalized for stochastic affine constraints in [50]. More recently, [51–53] have developed projection-free Frank-Wolfe approaches for problems with functional constraints. However, they assume smooth or structured nonsmooth objective functions.

1.1.1 Other projection-free

The predominant body of literature on projection-free methods, including the current papers, typically assumes the existence of a Linear Minimization Oracle (LMO) for the feasible set $\mathcal{X}$. However, recent alternative approaches in [54–64] utilize various techniques, such as separation Oracles, membership Oracles, Newton iterations, and radial dual transformations. It's worth noting that some of these oracles can be implemented using others, as demonstrated, for instance, in [56]. 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]).

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). This separation sets the stage for a unique *Lagrange multiplier update rule* of the form

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

Traditional Lagrange multiplier updates replace the right-hand-side with a maximum with 0, rather than a maximum with $[-h_i(y_{t+1})]_+$ (see, for example, the classic update rule for the dual subgradient algorithm in [1, 65, 66]). Our update is inspired by a related update used in [67] for a different class of problems. However, the update in [67] 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 [68, 69]. The objective function, describing utility and fairness [70, 71], can be nonsmooth due to piecewise linearities and/or the usage of $\min\{\cdot\}$.

Our algorithm also has potential in Quantum State Tomography (QST), which is a nonsmooth and stochastic problem. Earlier Frank-Wolfe-type work for this problem has employed smoothing techniques [72], or has focused on solving the maximum likelihood variant of QST [73].

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 [74–77]. Furthermore, the algorithm is adaptable to complex scenarios such as Fused Lasso regression [78], enforcing additional desired structures through functional constraints.

1.4 Notation

The set of positive 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 Euclidean 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 $\mathbb{R}^d$. 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 $\mathcal{X} \subseteq \mathbb{V}$, the problem is

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

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

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

where $h(x) = (h_1(x), \dots, h_m(x))^\top$ 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*

$$\max_{x,y} \{\|x - y\| : x, y \in \mathcal{X}\} \leq D.$$

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): For a given $v \in \mathbb{V}$, this oracle returns a random point $x \leftarrow \text{In-LMO}_{\mathcal{X}}\{v\}$ such that $x \in \mathcal{X}$ and*

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

with a known error bound $\delta > 0$.

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 $s, g_1, \dots, g_m$ such that*

$$\begin{aligned} &\mathbb{E}\{s \mid y\} \in \partial f(y), \\ &\mathbb{E}\{g_i \mid y\} \in \partial h_i(y) \quad \forall i \in \{1, \dots, m\}. \end{aligned}$$

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

$$\|g_i\| \leq G_i \quad \forall i \in \{1, \dots, m\}$$

$$\sum_{i=1}^{m} G_i^2 \leq G^2$$
$$\sqrt{\mathbb{E}\left\{\|s\|^2 \mid y\right\}} \leq L$$

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.*

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:*

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

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 original. 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:*

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

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

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

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:

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

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

The 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 [42] 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

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

$$= f^* + \langle \lambda \, , \, y - x \rangle + \langle \lambda \, , \, x - x^* \rangle$$
$$\overset{\text{(b)}}{\geq} f^* + \langle \lambda \, , \, y - x \rangle$$

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 $i \in \{1, \dots, m\}$. 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 [42]).

Taking the Euclidean norm from both sides of (10) and using the triangle inequality (note $\mu_i \geq 0$), we obtain:

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

Adding the Cauchy–Schwarz inequality, we get

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

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]):

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

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* [79] to overcome this. Part of this theorem demonstrates that if $r : \mathcal{Y} \to \mathbb{R}$ is a convex and ζ-Lipschitz continuous function on the convex set $\mathcal{Y}$, then there exists an extended function $\tilde{r} : \mathbb{V} \to \mathbb{R}$ that satisfies the following:

(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 $i \in \{1, \dots, m\}$. 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 $i \in \{1, \dots, m\}$) 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,

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

$$\left\| \tilde{f}'(x^*) \right\| \leq L,$$

which 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 noisy way) by the oracle In-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 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 example, if $\mathcal{Y}$ is a box, or a fixed radius ball centered at the origin, or the entire space $\mathbb{V}$ itself, then projections are trivial. 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. Our 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:

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

```
Require: Parameters: $T$, $\eta, \alpha, \beta$. Initial point: $x_1 \in \mathcal{X}$.
 1: $y_1 \leftarrow x_1$
 2: $Q_1 \leftarrow \mathbf{0}$
 3: Obtain stoch subgradients at $y_1$: $s_1, g_{1,1}, \ldots, g_{m,1}$
 4: $W_1 \leftarrow [-h(y_1)]_+$
 5: for $1 \leq t \leq T-1$ do
 6:     $x_{t+1} \leftarrow \text{In-LMO}_{\mathcal{X}}\{-Q_t\}$
 7:     $p_t \leftarrow \eta Q_t + s_t + \beta \sum_{i=1}^{m} (W_{i,t} + h_i(y_t)) g_{i,t}$
 8:     $\tilde{y}_{t+1} \leftarrow \dfrac{\left(\alpha + 2G^2\beta\right) y_t + \eta x_{t+1} - p_t}{\alpha + 2G^2\beta + \eta}$
 9:     $y_{t+1} \leftarrow \text{PO}_{\mathcal{Y}}\{\tilde{y}_{t+1}\}$
                        ▷ Lines 7,8 and 9 are only separated for better readability.
10:     $Q_{t+1} \leftarrow Q_t + y_{t+1} - x_{t+1}$
11:     Obtain stoch subgradients at $y_{t+1}$: $s_{t+1}, g_{1,t+1}, \ldots, g_{m,t+1}$
12:     for $1 \leq i \leq m$ do
13:         $W_{i,t+1} \leftarrow \max\left\{W_{i,t} + h_i(y_t) + \langle g_{i,t}\, ,\, y_{t+1} - y_t \rangle, [-h_i(y_{t+1})]_+\right\}$
14:     end for
15: end for
16: return $\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$
```

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

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

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

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

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.

Theorem 1 (Objective gap)**.** *Given Assumption 1-3, for Algorithm 1 with any $T \in \{1, 2, 3, ...\}$, $\eta > 0$, $\alpha > 0$, $\beta > 0$, the expected gap in the objective function is bounded as follows:*

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

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

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

while under Parameter Selection (ParSel.2) we have

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

Theorem 2 (Constraint violation)**.** *Given Assumption 1-3, Algorithm 1 under Parameter Selection (ParSel.1) yields*

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

while under Parameter Selection (ParSel.2) we have

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

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 1. *When the number of iterations T is on the order of $\mathcal{O}(\epsilon^{-2})$, the expected suboptimality $\mathbb{E}\{f(\bar{x})\} - f^*$ 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 all t and using Line 2 which states $Q_1 = 0$, gives:

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

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

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 $g_{i,t}$:

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

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

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

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

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

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

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

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 equations (15) and (16).

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

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

Using the iterated expectation gives:

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

Assumption 3.iii implies $\mathbb{E}\{g_{i,t} \mid y_t\} \in \partial h_i(y_t)$. Using equation (16) and convexity of the function h_i we get:

$$\mathbb{E}\left\{W_{i,t} + h_i(y_t) \mid y_t\right\}\left(h_i(y_t) + \langle \mathbb{E}\left\{g_{i,t} \mid y_t\right\}, x^* - y_t\rangle\right) \\ \leq \mathbb{E}\left\{W_{i,t} + h_i(y_t) \mid y_t\right\} h_i(x^*) \leq 0.$$

The last part is by the inequality $h_i(x^*) \leq 0$. This proves equation (17).

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

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

By summing over the range $1 \leq i \leq m$ and leveraging Assumption 3.iii, which implies $\sum_{i=1}^m \|g_{i,t}\|^2 \leq G^2$, and using the vectorized notation (14), we get the proof of equation (18).

□

3.2 Algorithm analysis

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

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

By iterated expectations we have

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

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

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

Taking expectations of both sides and rearranging gives

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

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

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

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 can be found in various forms in [80–82]. □

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 $x^* \in \mathcal{Y}$):

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

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:

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

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

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

$$+\frac{\beta}{2}\|W_{t+1}\|_2^2-\frac{\beta}{2}\|W_T\|_2^2-\frac{\beta}{2}\|\left[-h(y_{t+1})\right]_+\|_2^2+\frac{\beta}{2}\|h(y_t)\|_2^2$$

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

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

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

$$Q_1=\mathbf{0}$$
$$\|W_1\|_2=\|\left[-h(y_1)\right]_+\|_2\le\|h(y_1)\|_2$$
$$\|W_T\|_2\ge\|\left[-h(y_T)\right]_+\|_2$$

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), (23) becomes:

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

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

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

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

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

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

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

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

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

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

Consider the following:

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

Taking expectation we can simply write

$$\mathbb{E}\left\{\langle s_T \, , \, y_T - x^* \rangle\right\} \le \frac{\mathbb{E}\left\{\|s_T\|^2\right\}}{2\alpha} + \frac{\alpha}{2} \mathbb{E}\left\{\|x^* - y_T\|^2\right\} \quad (27)$$

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

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

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

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

$$\overset{\text{(c)}}{\geq} T\mathbb{E}\left\{f(\bar{x}) - L\|\bar{y}-\bar{x}\|\right\} - Tf(x^*)$$

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

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

The equation (13) states $Q_T = T\left(\bar{y} - \bar{x}\right)$. Thus by completing the square we can write:

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

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

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

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

□

4 A Numerical Experiment: Robust Reduced Rank Regression with Nuclear Norm Relaxation

The problem of multi-output regression [83], which is a special case of multi-task learning [84], 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:

$$y = cx + e.$$

Here, we defined matrices $y = (y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$. 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.

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 [74–76].

Reduced Rank Regression, introduced by Anderson in 1951 [85], is a specific form of multi-output regression. It operates under the assumption that the rank of the coefficient matrix c is small. In this approach, the relationship between multiple output variables and a set of input features is modeled using a lower-rank approximation. This allows us to capture underlying patterns in the data efficiently. However, the low-rank constraint doesn't define a convex set. To make this constraint convex, we can employ nuclear norm regularization. The nuclear norm encourages low-rank solutions by penalizing the sum of the singular values of c [86].

Thus, our optimization problem can be formulated as follows:

$$\text{minimize } \frac{1}{n}\sum_{k=1}^{n} \|y_i - cx_i\|_2 \text{ subject to } c \in \mathbb{R}^{q\times p}, \|c\|_* \leq \gamma. \tag{30}$$

Here, $\|\cdot\|_2$ denotes the Euclidean norm of a column vector. The choice of the $\|\cdot\|_2$ loss function, as opposed to $\|\cdot\|_2^2$ which is common in regression, increases robustness against outliers. Intuitively, this is more robust as the loss grows linearly instead of quadratically when distancing from the true value [87, 88].

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 simulated four algorithms, with three of them utilizing full SVD computation:

- Our new Algorithm with exact LMO.
- P-MOLES from [89].
- Projected subgradient descent (PGD).

Additionally, we executed our algorithm with inexact LMO, which involves a stochastic calculation of the largest eigenvalue using the Lanczos algorithm [12, 90, 91], to assess the computational advantages and trade-offs.

Figures 1 and 2 illustrate the expected loss relative to noiseless data for a fixed γ value of 350. Specifically, 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 Figure 2.

In Figure 3, we keep the number of iterations fixed at $T = 300$. When γ is very small, all four algorithms perform poorly. This is because the function class is too limited, and the models are overly simplified. As γ becomes very large, the performance of all models starts to deteriorate, which can be attributed to the broad scope of the model class. However, 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 in the results.[1]

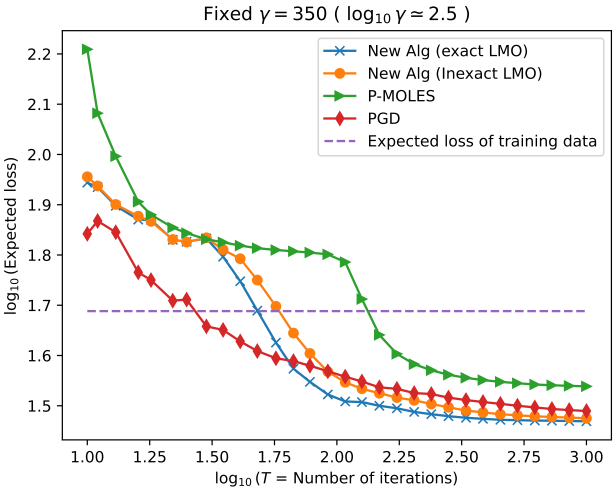

Fig. 1 Expected loss as a function of the number of iterations for $\gamma = 350$, compared to noiseless data.

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 [92, 93], an established method that substitutes the Euclidean 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 [94]. Another important area to explore is how our method works in online settings. This is especially relevant since projection-free online convex optimization is a highly discussed topic today [37, 57, 58, 61, 62, 95].

[1] It is worth mentioning that 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 the Frank-Wolfe-type algorithms. One possible idea to tackle this challenge is to restart the algorithm. This means that after achieving a low-error point, the algorithm is rerun, starting from that point.

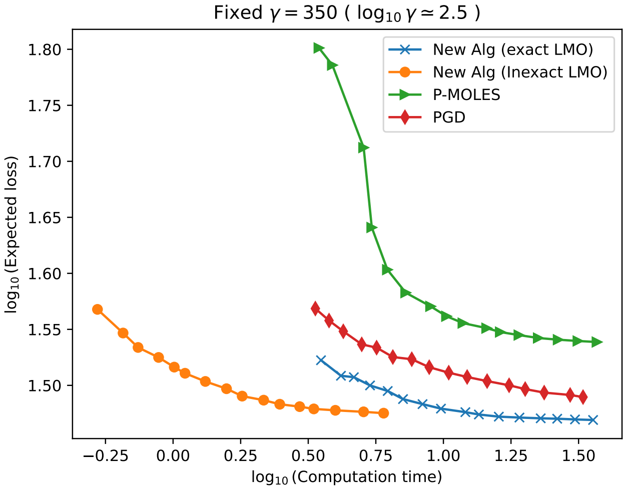

Fig. 2 Computational time versus error for inexact LMO with $\gamma = 350$, demonstrating the computational efficiency.

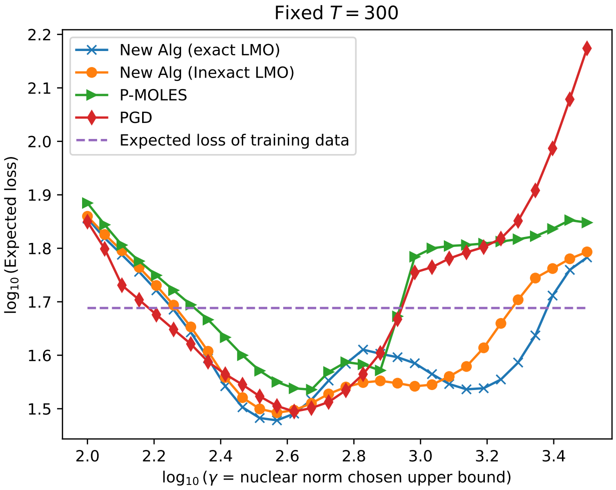

Fig. 3 Performance of four algorithms at $T = 300$ iterations. Performance dips for very small or large γ.

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, 59, 63, 64, 89, 96].

Declarations

Funding. This research received partial support from the National Science Foundation (NSF), including NSF CCF 1718477 and NSF SpecEES 1824418 grants.

Code and data availability. The data used in our simulation is synthetic. The codes and synthetically generated data are accessible at github.com/kamiarasgari.

Competing interests. 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, satisfy the following conditions:

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

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

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

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

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

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:

$$\mathbb{E}\{g_x \mid x\} \in \partial h_i(x), \quad \|g_x\| \leq G_i,$$

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:

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

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 $i \in \{1, \dots, m\}$ 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 equation (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 $\Omega \coloneqq \alpha + 2G^2\beta + \eta$. This allows us to further simplify it as:

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

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 equation (14), we conclude the proof. □

A.1 Proof of Theorem 2

Lemma 6. *: Line 13 implies the following inequality:*

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

Proof. : Fix $i \in \{1, \dots, m\}$. 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 $t \in \{1, \dots, T-1\}$ 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:

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

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

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

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

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

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:

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

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 the (A1) as an inequality for i-th element of vectors belonging to $\mathbb{R}_+^m$, we can write: [2]

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

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) in 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 equation (23) is demonstrated here:

[2] The vectors are as follows:

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

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

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:

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

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:

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

Using (27) in the above equation we get:

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

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

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

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:

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

Replacing this inequality in (A4) results in:

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

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

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

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

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

By using reverse triangle inequality we get

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

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

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

To connect this result to $\|\left[-h(y_T)\right]_+\|_2$, note that for any arbitrary vector $v \in \mathbb{R}^m$, the inequality $\|v\|_2^2 = \|\left[-v\right]_+\|_2^2 + \|\left[v\right]_+\|_2^2$ holds. Therefore, for $h(y_T) \in \mathbb{R}^m$, we can write:

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

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

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

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

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

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

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

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

$$\mathbf{LHS}_c = \frac{\beta}{2}\mathbb{E}\left\{\|W_T\|_2^2 + \|\left[h(y_T)\right]_+\|_2^2\right\} - \mathbb{E}\left\{\mu^\top h(y_T) + \mu^\top W_T\right\}$$

$$
\begin{aligned}
&\overset{(a)}{\geq} \frac{\beta}{4}\left(\mathbb{E}\left\{\|W_T\|_2 + \|\left[h(y_T)\right]_+\|_2\right\}\right)^2 - \mathbb{E}\left\{\mu^\top h(y_T) + \mu^\top W_T\right\} \\
&\overset{(b)}{\geq} \frac{\beta}{4}\left(\mathbb{E}\left\{\|W_T\|_2 + \|\left[h(y_T)\right]_+\|_2\right\}\right)^2 - \|\mu\|_2 \mathbb{E}\left\{\|\left[h(y_T)\right]_+\|_2 + \|W_T\|_2\right\} \\
&\overset{(c)}{=} \left(\frac{\sqrt{\beta}}{2}\mathbb{E}\left\{\|W_T\|_2 + \|\left[h(y_T)\right]_+\|_2\right\} - \frac{\|\mu\|_2}{\sqrt{\beta}}\right)^2 - \frac{\|\mu\|_2^2}{\beta}
\end{aligned}
$$

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

By employing these three inequalities, the equation (A10) can be rendered in a simplified form as follows:

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

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. This implies:

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

Which become

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

Substituting (A12) in Lemma 6 yields:

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

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 defined in (A11), we obtain:

$$\frac{\Gamma}{T} \leq \mathcal{O}\left(\epsilon\right),$$

and thus

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

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

$$
\begin{aligned}
\mathbb{E}\left\{\|\left[h(\bar{x})\right]_+\|_2\right\} &\overset{\text{(a)}}{\leq} \sqrt{\frac{\Gamma}{T}}\left(\left(1+\sqrt{2}\right)\sqrt{\frac{2}{\beta}}+\sqrt{\frac{2G^2}{T\eta}}\right)+\frac{3\|\mu\|_2}{\beta}+\frac{G\|\lambda\|}{T\eta} \\
&= \sqrt{\frac{\Gamma}{T}\left(1+\sqrt{2}\right)^2\frac{2}{\beta}}+\sqrt{\frac{\Gamma}{T}\frac{2G^2}{T\eta}}+\sqrt{\frac{9\|\mu\|_2^2}{\beta^2}}+\sqrt{\frac{G^2\|\lambda\|^2}{T^2\eta^2}} \\
&\overset{\text{(b)}}{\leq} \sqrt{\frac{\Gamma}{T}\left(\frac{47}{\beta}+\frac{8G^2}{T\eta}\right)+\frac{36\|\mu\|_2^2}{\beta^2}+\frac{4G^2\|\lambda\|^2}{T^2\eta^2}} \\
&\overset{\text{(c)}}{\leq} \sqrt{\frac{\Gamma}{T}\left(\frac{47}{\beta}+\frac{8G^2}{T\eta}\right)+\frac{36\|\mu\|_2^2}{\beta^2}+\frac{4G^2\left(L+\|\mu\|_2\,G\right)^2}{T^2\eta^2}}
\end{aligned}
\tag{A14}
$$

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

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

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:

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

and thus

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

where A_0, A_1 and A_2 are defined as follows:

$$
\begin{aligned}
A_0 \coloneqq& 47GLD^2+47GLD\sqrt{D^2+2\delta}+47G^2D^2+12G^2(D^2+2\delta) \\
&+8G^2D\sqrt{D^2+2\delta}+8\frac{G^3}{L}D\sqrt{D^2+2\delta}+\|h(x^*)\|_2^2\left(47+8\frac{G}{L}\frac{\sqrt{D^2+2\delta}}{D}\right)
\end{aligned}
$$

$$A_1 := 16\frac{G^3}{L}(D^2 + 2\delta) + 47G^2D\sqrt{D^2 + 2\delta}$$
$$A_2 := 55\frac{G^3}{L}D\sqrt{D^2 + 2\delta} + 83G^2D^2 + 8\frac{G^4}{L^2}(D^2 + 2\delta)$$

□

Remark 2. *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 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

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

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

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

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

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

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

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. Artificial Intelligence **306**, 103683 (2022)

[9] Combettes, C.W., Pokutta, S.: Complexity of linear minimization and projection on some sets. Operations Research Letters **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. Mathematical Programming **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 Research Logistics Quarterly **3**(1-2), 95–110 (1956) https://doi.org/10.1002/nav.3800030109 https://onlinelibrary.wiley.com/doi/pdf/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 Computational Mathematics and Mathematical Physics **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 Journal on Scientific Computing **38**(5), 3291–3317 (2016)

[24] Combettes, C.W., Pokutta, S.: Revisiting the approximate carathéodory problem via the frank-wolfe algorithm. Mathematical Programming **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. Mathematical Programming **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. Journal of Optimization Theory and Applications **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. Mathematical Programming **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. Advances in Neural Information Processing Systems **28** (2015)

[36] Duchi, J.C., Bartlett, P.L., Wainwright, M.J.: Randomized smoothing for stochastic optimization. SIAM Journal on Optimization **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. Advances in Neural Information Processing Systems **33**, 12211–12224 (2020)

[39] Polyak, V., Tret'yakov, N.: The method of penalty estimates for conditional extremum problems. USSR Computational Mathematics and Mathematical Physics **13**(1), 42–58 (1973)

[40] Lemaréchal, C., Nemirovskii, A., Nesterov, Y.: New variants of bundle methods. Mathematical programming **69**, 111–147 (1995)

[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] Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena Scientific, Raleigh (1999)

[43] Xu, Y.: Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. Mathematical Programming **185**, 199–244 (2021)

[44] Neely, M.J., Yu, H.: Lagrangian methods for o (1/t) convergence in constrained

convex programs. Convex Optimization: Theory, Methods, and Applications, edited by Arto Ruud, Nova Publishers (2019)

[45] Lan, G., Monteiro, R.D.: Iteration-complexity of first-order augmented lagrangian methods for convex programming. Mathematical Programming **155**(1-2), 511–547 (2016)

[46] Wei, X., Yu, H., Ling, Q., Neely, M.J.: Solving Non-smooth Constrained Programs with Lower Complexity than $\mathcal{O}(1/\varepsilon)$: A Primal-Dual Homotopy Smoothing Approach (2018)

[47] Yu, H., Neely, M.J.: A primal-dual type algorithm with the 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

[48] Yu, H., Neely, M., Wei, X.: Online convex optimization with stochastic constraints. Advances in Neural Information Processing Systems **30** (2017)

[49] Neely, M.J., Yu, H.: Online convex optimization with time-varying constraints. arXiv preprint arXiv:1702.04783 (2017)

[50] Wei, X., Neely, M.J.: Primal-Dual Frank-Wolfe for Constrained Stochastic Programs with Convex and Non-convex Objectives (2018)

[51] Lan, G., Romeijn, E., Zhou, Z.: Conditional gradient methods for convex optimization with general affine and nonlinear constraints. SIAM Journal on Optimization **31**(3), 2307–2339 (2021)

[52] Cheng, Y., Lan, G., Romeijn, H.E.: Functional constrained optimization for risk aversion and sparsity control. arXiv preprint arXiv:2210.05108 (2022)

[53] Lee, D., Ho-Nguyen, N., Lee, D.: Projection-Free Online Convex Optimization with Stochastic Constraints (2023)

[54] Mahdavi, M., Yang, T., Jin, R., Zhu, S., Yi, J.: Stochastic gradient descent with only one projection. Advances in neural information processing systems **25** (2012)

[55] 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

[56] Lee, Y.T., Sidford, A., Vempala, S.S.: Efficient convex optimization with membership oracles. In: Conference On Learning Theory, pp. 1292–1294 (2018). PMLR

[57] Mhammedi, Z.: Efficient projection-free online convex optimization with membership oracle. In: Conference on Learning Theory, pp. 5314–5390 (2022).

PMLR

[58] 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

[59] 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

[60] Garber, D., Kretzu, B.: New Projection-free Algorithms for Online Convex Optimization with Adaptive Regret Guarantees (2023)

[61] Gatmiry, K., Mhammedi, Z.: Projection-Free Online Convex Optimization via Efficient Newton Iterations (2023)

[62] Garber, D., Kretzu, B.: Projection-free online exp-concave optimization. arXiv preprint arXiv:2302.04859 (2023)

[63] Grimmer, B.: Radial Duality Part II: Applications and Algorithms (2022)

[64] Grimmer, B.: Radial Duality Part I: Foundations (2023)

[65] Bertsekas, D.: Convex Optimization Theory vol. 1. Athena Scientific, Raleigh (2009)

[66] Bertsekas, D.: Convex Optimization Algorithms. Athena Scientific, Raleigh (2015)

[67] Yu, H., Neely, M.J.: A primal-dual parallel method with $\mathcal{O}(1/\epsilon)$ convergence for constrained composite convex programs. arXiv preprint arXiv:1708.00322 (2017)

[68] Kelly, F.P., Maulloo, A.K., Tan, D.K.H.: Rate control for communication networks: shadow prices, proportional fairness and stability. Journal of the Operational Research society **49**(3), 237–252 (1998)

[69] Bertsekas, D.P.: Linear Network Optimization: Algorithms and Codes. MIT press, Cambridge (1991)

[70] Bertsekas, D.P.: Network Optimization Continuous and Discrete Models. Athena Scientific, Raleigh (1998)

[71] Neely, M.J.: Stochastic network optimization with application to communication and queueing systems. Synthesis Lectures on Communication Networks **3**(1), 1–211 (2010)

[72] Hazan, E.: Sparse approximate solutions to semidefinite programs. In: Latin American Symposium on Theoretical Informatics, pp. 306–316 (2008). Springer

[73] Shinde, N., Narayanan, V., Saunderson, J.: An inexact frank-wolfe algorithm for composite convex optimization involving a self-concordant function. arXiv preprint arXiv:2310.14482 (2023)

[74] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust statistics. Wiley series in probability and statistics (2005)

[75] 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

[76] Lerasle, M.: Selected topics on robust statistical learning theory. Lecture Notes (2019)

[77] Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**, 273–297 (1995)

[78] 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

[79] Petrakis, I.: McShane-Whitney extensions in constructive analysis. Logical Methods in Computer Science **Volume 16 Issue 1** (2020) https://doi.org/10.23638/LMCS-16(1:18)2020

[80] Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using bregman functions. SIAM Journal on Optimization **3**(3), 538–543 (1993)

[81] Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization **19**(4), 1574–1609 (2009)

[82] Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. submitted to SIAM Journal on Optimization **2**(3) (2008)

[83] 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)

[84] Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering **34**(12), 5586–5609 (2021)

[85] Anderson, T.W.: Estimating linear restrictions on regression coefficients for multivariate normal distributions. The Annals of Mathematical Statistics, 327–351 (1951)

[86] Chen, K., Dong, H., Chan, K.-S.: Reduced rank regression via adaptive nuclear

norm penalization. Biometrika **100**(4), 901–920 (2013)

[87] 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)

[88] 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)

[89] Thekumparampil, K.K.: Optimal nonsmooth frank-wolfe method for stochastic regret minimization. (2020). https://api.semanticscholar.org/CorpusID:229347497

[90] Lanczos, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators (1950)

[91] Kuczyński, J., Woźniakowski, H.: Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. SIAM Journal on Matrix Analysis and Applications **13**(4), 1094–1122 (1992) https://doi.org/10.1137/0613066 https://doi.org/10.1137/0613066

[92] 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 https://doi.org/10.1137/1027074

[93] Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters **31**(3), 167–175 (2003)

[94] Bach, F., Moulines, E.: Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n) (2013)

[95] Hazan, E., Minasyan, E.: Faster projection-free online learning. In: Conference on Learning Theory, pp. 1877–1893 (2020). PMLR

[96] 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