SIAM J. OPTIM.
Vol. 22, No. 2, pp. 702–727

DOUBLE SMOOTHING TECHNIQUE FOR LARGE-SCALE LINEARLY CONSTRAINED CONVEX OPTIMIZATION*

OLIVIER DEVOLDER†, FRANÇOIS GLINEUR†, AND YURII NESTEROV†

Abstract. In this paper, we propose an efficient approach for solving a class of large-scale convex optimization problems. The problem we consider is the minimization of a convex function over a simple (possibly infinite-dimensional) convex set, under the additional constraint $\mathcal{A}u \in T$, where $\mathcal{A}$ is a linear operator and T is a convex set whose dimension is small compared to the dimension of the feasible region. In our approach, we dualize the linear constraints, solve the resulting dual problem with a purely dual gradient-type method and show how to reconstruct an approximate primal solution. Because the linear constraints have been dualized, the dual objective function typically becomes separable, and therefore easy to compute. In order to accelerate our scheme, we introduce a novel double smoothing technique that involves regularization of the dual problem to allow the use of a fast gradient method. As a result, we obtain a method with complexity $O(\frac{1}{\epsilon}\ln\frac{1}{\epsilon})$ gradient iterations, where ϵ is the desired accuracy for the primal-dual solution. Our approach covers, in particular, optimal control problems with a trajectory governed by a system of linear differential equations, where the additional constraints can, for example, force the trajectory to visit some convex sets at certain moments in time.

Key words. convex optimization, fast gradient methods, infinite-dimensional, Hilbert space, complexity theory, smoothing technique, optimal control

AMS subject classifications. 90C06, 90C25, 90C60

DOI. 10.1137/110826102

1. Introduction.

In large-scale convex optimization, first-order methods are the methods of choice due to their cheap iteration cost. In particular, constrained problems can be solved provided that performing projection on their feasible set is computationally easy. In this work, we assume that both the convex objective function J, defined on the Hilbert space U, and the convex feasible region $S \subset U$ are sufficiently simple so that the problem $\min_{u\in S} J(u)$ can be solved efficiently, or even in closed form. However, the situation becomes completely different when adding the constraint $\mathcal{A}u \in T$, based on a linear operator $\mathcal{A} : U \to V^*$, where V is another Hilbert space and T is a bounded closed convex set in V^* (the dual space of V). Indeed, the problem may become difficult because projection onto the new feasible set $\{u \in S : \mathcal{A}u \in T\}$ may be computationally very expensive, or even intractable.

A natural approach is, therefore, to dualize this difficult linear constraint, obtaining the primal-dual pair of problems

*Received by the editors February 28, 2011; accepted for publication (in revised form) April 11, 2012; published electronically June 26, 2012.
http://www.siam.org/journals/siopt/22-2/82610.html

†Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), B-1348 Louvain-la-Neuve, Belgium; and Université catholique de Louvain, Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM) Institute, B-1348 Louvain-la-Neuve, Belgium (Olivier.Devolder@uclouvain.be, Francois.Glineur@uclouvain.be, Yurii.Nesterov@uclouvain.be). The first author is an F.R.S.-FNRS Research Fellow. This text presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The third author acknowledges support from Office of Naval Research grant N000140811104: Efficiently Computable Compressed Sensing. The scientific responsibility rests with its authors.

$$P^* = \min_{u\in S}\left\{J(u) + \max_{z\in V}[\langle \mathcal{A}u, z\rangle - \sigma_T(z)]\right\},$$
$$D^* = \max_{z\in V}\left\{-\sigma_T(z) + \min_{u\in S}[J(u) + \langle \mathcal{A}u, z\rangle]\right\},$$

where $\sigma_T(z) = \sup_{x\in T}\langle x, z\rangle$ denotes the support function of set T, defined on V.

In this paper, we assume that the dimension of set V (i.e., the size of the linear constraints) is small compared to the dimension of set U, the latter being allowed to be infinite. Due to this asymmetry, we are led to consider a purely dual algorithmic scheme, generating its iterates only in the low-dimensional space V. The only operation we need to be able to perform in the infinite-dimensional or high-dimensional space U is the computation of the value of dual objective function at a given point $z \in V$, which requires solving the optimization subproblem $\min_{u\in S} J(u) + \langle \mathcal{A}u, z\rangle$ over the simple set S.

Example 1. As a first motivation, consider the purely linear infinite-dimensional problem

$$P^* = \inf_{u\in L^2([\alpha,\beta]):M_1\le u(t)\le M_2} \int_\alpha^\beta f(t)u(t)dt : \quad \int_\alpha^\beta a_i(t)u(t)dt = b_i \quad \forall i = 1,\dots,m,$$

where we seek a function u in Hilbert space $L^2([\alpha,\beta])$, and data consists of functions a_i ($1 \le i \le m$) and f in $L^2([\alpha,\beta])$. Since we are working in L^2, inequalities on $u(t)$ are to be understood as to hold almost everywhere. It is straightforward to define

$$U = L^2([\alpha,\beta]), \quad S = \{u \in U : M_1 \le u(t) \le M_2, t\in[\alpha,\beta]\}, \quad J(u) = \int_\alpha^\beta f(t)u(t)dt,$$
$$V = R^m, \quad T = \{b\} \subset V, \quad \mathcal{A} : U \to V^* : u \to \left(\int_\alpha^\beta a_1(t)u(t)dt, \dots, \int_\alpha^\beta a_m(t)u(t)dt\right)^T$$

so that the problem fits our formulation $\min_{u\in S} J(u) : \mathcal{A}u \in T$. Dualizing the linear equality constraints, we obtain the dual function

$$\Theta(z) = -\sum_{i=1}^m z_i b_i + \inf_{u\in S}\left[\int_\alpha^\beta f(t)u(t)dt + \sum_{i=1}^m \left(\int_\alpha^\beta a_i(t)u(t)dt\right) z_i\right] \tag{1.1}$$
$$= -\sum_{i=1}^m z_i b_i + \inf_{u\in S}\int_\alpha^\beta \left(f(t) + \sum_{i=1}^m a_i(t)z_i\right) u(t)dt. \tag{1.2}$$

Due to the fact that only pointwise constraints $M_1 \le u(t) \le M_2$ are still present in this problem, we can solve it in a pointwise way, minimizing u for each value of t separately. Indeed, any solution $u_z \in S$ satisfying $u_z(t) = M_1$ when $f(t) + \sum_{i=1}^m a_i(t)z_i > 0$ and $u_z(t) = M_2$ when $f(t) + \sum_{i=1}^m a_i(t)z_i < 0$ is the optimal solution for problem (1.1).

Since we are able to compute the value of $\Theta(z)$ in closed form for any value of z, we can apply a first-order method to the finite-dimensional problem $\max_{z\in V}\Theta(z)$.

Our goal in this work is to show that it is possible to solve the dual problem efficiently and reconstruct from this process a nearly optimal and feasible primal solution. We develop to that effect a new double smoothing approach, which is a variant of the smoothing techniques described in [20, 21, 22]. This technique uses the problem structure to regularize the dual objective function into a smooth strongly convex function with Lipschitz continuous gradient. These modifications allow us to minimize the dual function with an optimal gradient scheme in $O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)$iterations, where ϵ is the desired accuracy. From the dual minimization sequence, we reconstruct a nearly feasible and optimal primal solution, whose accuracy can be controlled by parameters of our algorithm.

The structure of this paper is as follows. In section 2, we recall briefly two standard approaches for solving a nonsmooth convex optimization problem with a first-order method: subgradient-type schemes and smoothing techniques. We also present the first-order methods that can be used to solve efficiently the smoothed problem obtained by the smoothing technique. In particular, we recall the optimal method [19] for smooth and strongly convex functions and describe its rate of convergence. In section 3, we present in a more general form our problem class and derive the corresponding dual problem. Using Danskin's theorem, we show that the dual objective function is, in general, nonsmooth. Section 4 presents two simple examples of problems (one finite-dimensional, the other infinite-dimensional) with separable structure that fit our problem class. The double smoothing is described in section 5, where we apply two regularizations to the dual objective function in order to make it smooth and strongly convex. We also explain the necessity of requiring both properties. In section 6, we study under which regularity conditions strong duality holds and how it is possible to bound the size of the dual optimal set. This bound will be useful in the convergence analysis of our scheme. In section 7, an optimal first-order method is applied to the modified dual objective function and a nearly feasible and optimal primal solution is reconstructed from the dual minimization sequence. Accuracy of the primal and dual solutions can be adjusted by parameters of our algorithm. In section 8, we discuss the practical implementation of our method; in particular, when the size of the dual optimal set is unknown. In section 9, we consider applications of our double smoothing technique to optimal control problems. We conclude this paper with a comparison between our results and the existing literature.

2. First-order methods in convex optimization.

Consider the convex optimization problem $\min_{y\in V} f(y)$, where $f : V \to R$ is a convex function defined on the finite-dimensional space V. If f is nondifferentiable, we know that the complexity of a black-box first-order method that does not use the problem structure cannot be better that $O\left(\frac{1}{\epsilon^2}\right)$ iterations, where ϵ is the desired accuracy for the objective function (see [18, 19]). This lower bound is achievable by various first-order methods for nonsmooth convex problems, such as subgradient methods (see, e.g., [19, 23]). These schemes can, therefore, be applied directly to a nonsmooth convex function, albeit with a relatively slow convergence rate.

When the nonsmooth function has a particular saddle-point structure, then

$$f(y) = \max_{u\in S}\{g(u) + \langle Au, y\rangle\}, \tag{2.1}$$

where $g : U \to R$ is concave on the finite-dimensional space U and $S \subset U$ is closed and convex, another approach can be used. In the smoothing technique developed in [20, 21, 22], this nonsmooth function is approximated by a smooth one and an optimal first-order method of smooth convex optimization is applied to the smooth approximation. With this approach, we can solve the original nonsmooth problem up to accuracy ϵ in only $O\left(\frac{1}{\epsilon}\right)$ iterations (instead of $O\left(\frac{1}{\epsilon^2}\right)$ with subgradient scheme). We follow this smoothing approach in this paper, and the efficiency of the first-order method of smooth convex optimization used for minimizing the smoothed approximation therefore plays an important role.

In smooth convex optimization, an important class of objective function is $F_L^{1,1}(V)$, the class of convex function $f : V \to R$ with Lipschitz-continuous gradient, i.e.,

$$\|\nabla f(x) - \nabla f(y)\|_{V^*} \leq L \|x - y\|_V \quad \forall x, y \in V$$

for some $L > 0$. The easiest and most classical numerical scheme that can be applied to such a problem is the gradient method. However, it is well known that this method exhibits nonoptimal complexity of $O(\frac{L}{\epsilon})$ iterations where ϵ is the desired accuracy for the objective function. Several variants of first-order methods for the class $F_L^{1,1}(Q)$ that achieve the lower complexity bound of $O(\sqrt{\frac{L}{\epsilon}})$ iterations have been known since 1983 [24, 25, 19, 20]. These schemes, also called fast gradient methods, outperform theoretically and very often in practice the classical gradient method.

In the smoothing approach, we want to minimize the original nonsmooth function with an accuracy ϵ. We construct a smooth approximation of f belonging to $F_{L(\epsilon)}^{1,1}$ with constant $L(\epsilon) = \Theta(\frac{1}{\epsilon})$. Applying a fast gradient method for $F_{L(\epsilon)}^{1,1}$ to this smoothed function, we can solve the original problem with the desired accuracy in $O(\frac{1}{\epsilon})$ iterations. In our work, the function that we optimize using a first-order method, with saddle-point structure (2.1), is the dual objective function. However, our goal is to solve efficiently the primal problem, not the dual one.

In order to reconstruct a good primal solution from the iterates of the numerical scheme applied to the dual problem, we will need to apply a second smoothing to the dual function before we apply the fast gradient method. Its purpose will be to ensure strong convexity of the resulting dual objective function. Let $S_{\kappa,L}^{1,1}(V)$ be the class of functions $f \in F_L^{1,1}(V)$ which are strongly convex with parameter $\kappa > 0$. Fast gradient methods that are optimal for $S_{\kappa,L}^{1,1}(V)$ are also known (see, for example, [19]), and we will use such a method to minimize the doubly smoothed dual objective function.

For the reader's convenience, we conclude this section with a presentation of the simplest optimal method for minimizing smooth strongly convex functions. Let function $f \in S_{\kappa,L}^{1,1}(V)$ and consider the problem $\min_{y \in V} f(y)$. We assume that this problem is solvable. Denote by f^* its optimal value and by y^* the optimal solution.

Algorithm. (see [19]). Choose $w_0 = y_0 \in V$.

(2.2) *Iteration* ($k \geq 0$): Set $y_{k+1} = w_k - \frac{1}{L}\nabla f(w_k)$, and compute
$$w_{k+1} = y_{k+1} + \frac{\sqrt{L}-\sqrt{\kappa}}{\sqrt{L}+\sqrt{\kappa}}(y_{k+1} - y_k).$$

By Theorem 2.2.3 in [19] we have

$$f(y_k) - f^* \leq \left(f(y_0) - f^* + \frac{\kappa}{2}\|y_0 - y^*\|_V^2\right) e^{-k\sqrt{\frac{\kappa}{L}}} \leq 2(f(y_0) - f^*)e^{-k\sqrt{\frac{\kappa}{L}}}. \tag{2.3}$$

Since ∇f is Lipschitz-continuous, in view of Theorem 2.1.5 in [19] we have

$$\frac{1}{2L}\|\nabla f(y_k)\|_{V^*}^2 \leq f(y_k) - f^* \overset{(2.3)}{\leq} 2(f(y_0) - f^*)e^{-k\sqrt{\frac{\kappa}{L}}}.$$

Therefore,

$$\|\nabla f(y_k)\|_{V^*}^2 \leq 4L(f(y_0) - f^*)e^{-k\sqrt{\frac{\kappa}{L}}}. \tag{2.4}$$

Finally, since f is strongly convex, by Theorem 2.1.8 in [19] we have

$$\frac{\kappa}{2}\|y_k - y^*\|_V^2 \leq f(y_k) - f^* \overset{(2.3)}{\leq} 2(f(y_0) - f^*)e^{-k\sqrt{\frac{\kappa}{L}}}.$$

Using this inequality and additional arguments, we conclude that

$$\|y_k - y^*\|_V^2 \leq \min\left\{\|y_0 - y^*\|_V^2, \frac{4}{\kappa}(f(y_0) - f^*)e^{-k\sqrt{\frac{\kappa}{L}}}\right\}. \tag{2.5}$$

3. Problem formulation and dual approach. As described in the introduction, we consider in this work optimization problems of the form

$$P^* = \inf_{u \in S} J(u) : \mathcal{A}u \in T, \tag{3.1}$$

where U is a Hilbert space endowed with the Euclidean norm $\|\cdot\|_U = \sqrt{(\cdot|\cdot)_U}$, S is a bounded, closed, convex set in U, V is a finite-dimensional Hilbert space endowed with the Euclidean norm $\|\cdot\|_V = \sqrt{(\cdot|\cdot)_V}$, T is a bounded, closed, convex set in V^*, the dual space of V, $J : U \to R$ is a closed and convex functional defined on S, and $\mathcal{A} : U \to V^*$ is a bounded linear operator. Space U is allowed to be infinite-dimensional, but the approach used in this paper is also efficient for large-scale finite-dimensional problems, i.e., when $\dim U \gg \dim V$.

Remark 1. Note that problems with multiple linear constraints also belong to problem class (3.1):

$$P^* = \inf_{u \in S} J(u) : \mathcal{A}_i u \in T_i \quad \forall i = 1, \ldots, m. \tag{3.2}$$

Indeed, assume that $V = V_1 \times V_2 \times \cdots \times V_m$, where V_i is a finite-dimensional Hilbert space for $i = 1, \ldots, m$. For each i, $\mathcal{A}_i : U \to V_i^*$ is a bounded linear operator. Let $T = T_1 \times T_2 \times \cdots \times T_m$, where T_i is a bounded, closed, convex set in V_i^*. Defining the linear operator $\mathcal{A} : U \to V^*$ such that $\langle \mathcal{A}u, z\rangle = \sum_{i=1}^m \langle \mathcal{A}_i u, z_i\rangle$ for every $u \in U$ and every $z = (z_1, \ldots, z_m) \in V$, we have the constraint $\mathcal{A}u \in T$ is clearly equivalent to $\mathcal{A}_i u \in T_i \forall i = 1, \ldots, m$. Finally, we have

$$\begin{aligned} \|\mathcal{A}\| &= \max_{\|u\|_U = 1, \|z\|_V = 1} \langle \mathcal{A}u, z\rangle = \max_{\|u\|_U = 1, \sum_{i=1}^m \|z_i\|_{V_i}^2 = 1} \sum_{i=1}^m \langle \mathcal{A}_i u, z_i\rangle \\ &= \max_{\|u\|_U = 1} \left[\sum_{i=1}^m \|\mathcal{A}_i u\|_{V_i^*}^2\right]^{1/2} \leq \left[\sum_{i=1}^m \|\mathcal{A}_i\|^2\right]^{1/2}. \end{aligned}$$

Our assumptions on J, S, and T are motivated by the following classical result (see, for example, [26]).

THEOREM 3.1. *If X is a reflexive Banach space, $M \subset X$ is a bounded, closed, convex set and $F : X \to R$ is a closed, convex function, then an optimization problem $\min_{x \in M} F(x)$ is solvable. Furthermore, if, in addition, X is an Hilbert space and F is strongly convex, then the optimal solution of this problem is unique.*

We conclude that subproblems of the forms

$$\inf_{u \in S}\{J(u) + \langle \mathcal{A}u, z\rangle\} \quad \text{and} \quad \inf_{x \in T} \langle x, z\rangle$$

are solvable for each $z \in V$, and that subproblems of the form

$$\inf_{u \in S}\left\{J(u) + \langle \mathcal{A}u, z\rangle + \frac{\mu}{2}\|u\|_U^2\right\} \quad \text{and} \quad \inf_{x \in T}\left\{\langle x, z\rangle + \frac{\rho}{2}\|x\|_{V^*}^2\right\}$$

have a unique optimal solution for every $z \in V$, $\mu > 0$ and $\rho > 0$.

In our setting, it is natural to dualize the linear constraint $\mathcal{A}u \in T$ and to consider a dual method, working only in the small-dimensional space V. Since T is a closed convex set, inclusion $\mathcal{A}u \in T$ is equivalent to $\langle \mathcal{A}u, z\rangle \leq \sigma_T(z)\ \forall z \in V$, where $\sigma_T(z) = \sup_{x\in T}\langle x, z\rangle$ denotes the support function of T, which allows us to dualize the linear constraints. We obtain the primal-dual pair of problems

$$P^* = \inf_{u\in S}\left[J(u) + \sup_{z\in V}(\langle \mathcal{A}u, z\rangle - \sigma_T(z))\right], \quad D^* = \sup_{z\in V}\left[-\sigma_T(z) + \inf_{u\in S}(J(u) + \langle \mathcal{A}u, z\rangle)\right].$$

Thus, the Lagrangian dual problem (in minimization form) is given by

$$-D^* = \theta^* = \inf_{z\in V}[\sigma_T(z) + \phi(z)] := \inf_{z\in V}\theta(z) \geq -P^*, \tag{3.3}$$

where we define $\phi(z) = \sup_{u\in S}[-J(u) - \langle \mathcal{A}u, z\rangle]$, we recall $\sigma_T(z) = \sup_{x\in T}\langle x, z\rangle$, and we let $\theta(z) = \sigma_T(z) + \phi(z)$.

We made the initial assumption that it is easy to optimize a function over simple set S, so that it is easy to compute the value of $\theta(z)$ for every $z \in V$; however, this function is typically nondifferentiable. Indeed, using Danskin's theorem [11, 3], the subdifferentials of σ_T and ϕ are given by

$$\partial\sigma_T(z) = \{x \in T : \langle x, z\rangle = \sigma_T(z)\},$$
$$\partial\phi(z) = \{-\mathcal{A}u : -J(u) - \langle \mathcal{A}u, z\rangle = \phi(z), u \in T\}.$$

As the optimization problems defining $\sigma_T(z)$ and $\phi(z)$ can have multiple optimal solutions, the above subdifferentials may contain several elements and function $\theta(z)$ can be nonsmooth. Dualization of problem (3.1), therefore, results in a nonsmooth convex problem.

As explained in the previous section, instead of relying on subgradient-type schemes with relatively slow convergence, we will solve the dual problem with a smoothing technique [20, 21, 22]. In the smoothing approach, using the specific structure of the problem, we apply some regularization to the objective function and obtain much faster methods (which are not pure black-box schemes anymore). We develop in this paper an algorithm capable of solving the dual problem with accuracy ϵ and to reconstruct, from a nearly optimal dual solution, a nearly optimal and feasible primal solution in $O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)$ iterations.

4. Examples.

Before we go into the details of the double smoothing, we provide two examples of problems with separable structure (one finite-dimensional, the other infinite-dimensional) that belong to our problems class.

4.1. A finite-dimensional example.

Consider the case where

- $U = R^N = R^{N_1} \times R^{N_2} \times \cdots \times R^{N_m}$ and $S = S_1 \times S_2 \times \cdots \times S_m$, where $S_i \subset R^{N_i}$,
- $V = R^n$ with $n \ll N$ and T is a bounded, closed, convex set in R^n,
- $J(u) = \sum_{i=1}^m J_i(u_i)$, where $u_i \in R^{N_i}$ and $J_i : R^{N_i} \to R$ is a closed and convex function,
- $\mathcal{A} = (A_1\ A_2\ \cdots\ A_m) \in R^{n\times N}$, where $A_i \in R^{n\times N_i}$.

Our problem becomes $\min \sum_{i=1}^m J_i(u_i) : \sum_{i=1}^m A_i u_i \in T, u_i \in S_i\ \forall i = 1, \ldots, m$. This problem has a specific structure that we want to exploit. When the coupling constraint $\sum_{i=1}^m A_i u_i \in T$ is dropped, we obtain a separable problem that we can solve separately for each u_i. With this property, it seems natural to dualize the coupling constraint and to consider the dual problem: $\min_{z\in R^n}\theta(z)=\min_{z\in R^n}\sigma_T(z)+\phi(z)$ in the small-dimensional space R^n. For each $z\in R^n$, the dual objective function can be computed in a pointwise way: solving the subproblem $\max_{u\in S}\{-J(u)-\langle Au,z\rangle\}$ is equivalent to solving separately for each i the subproblems $\max_{u_i\in S_i}\{-J_i(u_i)-\langle A_iu_i,z\rangle\}$, which we we assume to be computationally easy. The modified dual objective function obtained after smoothing $\max_{u\in S}\{-J(u)-\langle \mathcal{A}u,z\rangle-\frac{\mu}{2}\|u\|^2_{R^N}\}$ is similarly easy to compute with independent subproblems, since squared Euclidean norm $\|u\|_{R^N}$ is also separable: $\|u\|^2_{R^N}=\sum_{i=1}^m\|u_i\|^2_{R^{N_i}}$.

4.2. An infinite-dimensional example. Consider the case where

- $U=L^2([0,\mathcal{T}],R^m)$ and $S=\{u\in U: u(t)\in S(t)\ \forall t\in[0,\mathcal{T}]\}$, where $S(t)$ is a closed, convex set in R^m for each $t\in[0,\mathcal{T}]$ and $\cup_{t\in[0,\mathcal{T}]}S(t)$ is bounded,
- $V=R^n$ and T is a bounded, closed, convex set in R^n,
- $J(u)=\int_0^{\mathcal{T}}F(t,u(t))dt$, where function $F:[0,\mathcal{T}]\times R^m\to R$ is convex and continuous,
- $\mathcal{A}:U\to R^n$ is defined by $\mathcal{A}u=\int_0^{\mathcal{T}}A(t)u(t)dt$, where $\int_0^{\mathcal{T}}\|A(t)\|_2^2\,dt<+\infty$ and $A(t)\in R^{n\times m}\ \forall t\in[0,\mathcal{T}]$.

Our problem becomes

$$\inf\int_0^{\mathcal{T}}F(t,u(t))dt:\int_0^{\mathcal{T}}A(t)u(t)dt\in T, u(t)\in S(t)\quad\forall t\in[0,\mathcal{T}]. \tag{4.1}$$

When the linear coupling constraint is dropped, we obtain a separable problem that we can solve independently for each $t\in[0,\mathcal{T}]$: $\min_{u(t)\in S(t)}F(t,u(t))$. Hence, dualization of the linear coupling constraint here is also a natural approach. For each $z\in R^m$, the dual objective function can be computed in a pointwise way. Indeed, solving the maximization problem involved in $\phi(z)$,

$$\phi(z)=\sup_{u\in S}\{-J(u)-\langle\mathcal{A}u,z\rangle\}=\sup_{u(t)\in S(t)\ \forall t\in[0,\mathcal{T}]}-\int_0^{\mathcal{T}}F(t,u(t))-\langle A(t)u(t),z\rangle dt$$

is equivalent to solving independently for each value of $t\in[0,\mathcal{T}]$, a subproblem over $S(t)\subset R^m$: $\max_{v\in S(t)}[-F(t,v)-\langle v,A(t)^Tz\rangle]$, which is assumed to be easy to solve or even computable in closed form. The same separability property holds after smoothing, since minimizing the smoothed dual objective function $\sup_{u\in S}\{-J(u)-\langle\mathcal{A}u,z\rangle-\frac{\mu}{2}\|u\|_U^2\}$ is equivalent to the solving pointwise subproblems $\max_{v\in S(t)}\{-F(t,v)-\langle v,A(t)^Tz\rangle-\frac{\mu}{2}\|v\|^2_{R^m}\}$.

5. Double smoothing technique. We propose to solve the dual problem (3.3) using a new primal-dual smoothing technique. Note that, as shown above, its objective function is, in general, not differentiable and not strongly convex. However, we can ensure these properties by a double primal-dual regularization of θ. The goal of the first regularization is to obtain an objective function with Lipschitz-continuous gradient, for which we can apply much more efficient algorithms of smooth convex optimization. The goal of the second regularization is to obtain a strongly convex dual objective. This property is necessary to allow us to reconstruct efficiently a nearly feasible and optimal primal solution from a nearly optimal dual solutions.

5.1. First smoothing. Let us start from ensuring smoothness of the dual function. Dual objective $\theta(z)$ is a sum of two functions, both of which can be nonsmooth. Their nonsmoothness comes from the fact that the optimization problems defining

$\sigma_T(z)$ and $\phi(z)$ can have multiple optimal solutions at a given point z. A natural way to obtain a smooth approximation of θ is to modify these optimization subproblems in order to ensure the uniqueness of optimal solutions for each $z \in V$. For any $\rho > 0$, we can approximate $\sigma_T(z) = \sup_{x\in T}\langle x, z\rangle$ by a modified function

$$\sigma_{\rho,T}(z) = \sup_{x\in T}\left\{\langle x, z\rangle - \frac{\rho}{2}\|x\|_{V^*}^2\right\}. \tag{5.1}$$

In the same way, for any $\mu > 0$, we modify function $\phi(z)$ as follows:

$$\phi_\mu(z) = \sup_{u\in S}\left\{-J(u) - \langle \mathcal{A}u, z\rangle - \frac{\mu}{2}\|u\|_U^2\right\}. \tag{5.2}$$

The following result can be seen as an easy generalization of Theorem 1 in [20].

THEOREM 5.1. *Let H_1, H_2 be two Hilbert spaces. Assume that the linear operator $A : H_1 \to H_2^*$ is bounded and that the function $G : H_1 \to R$ is closed and strongly convex with parameter κ. Let $Q \subset \operatorname{dom}(G)$ be a closed, convex set. Then the function*

$$F(z) = \sup_{u\in Q}\{-G(u) - \langle \mathcal{A}u, z\rangle\} \tag{5.3}$$

is smooth with Lipschitz-continuous gradient $\nabla F(z) = -\mathcal{A}u_z$, where u_z is the unique optimal solution of the optimization problem defining $F(z)$. The Lipschitz constant of the gradient is equal to $\frac{\|A\|^2}{\kappa}$ where $\|A\| = \sup\{\langle \mathcal{A}u, z\rangle : u \in H_1, \|u\|_{H_1} = 1, z \in H_2, \|z\|_{H_2} = 1\}$.

Choosing now $H_1 = U$, $Q = S$, $H_2 = V$, $A = \mathcal{A}$, and $G(u) = J(u) + \frac{\mu}{2}\|u\|_U^2$, we conclude that ϕ_μ is smooth and convex with a Lipschitz continuous gradient equal to $\nabla\phi_\mu(z) = -\mathcal{A}u_{\mu,z}$, where $u_{\mu,z}$ denotes the unique optimal solution of the problem (5.2). Its Lipschitz constant is given by $L(\phi_\mu) = \frac{\|\mathcal{A}\|^2}{\mu}$. Similarly, if we choose $H_1 = V^*$, $Q = T$, $H_2 = V$, $A = I : V^* \to V^*$, and $G(x) = \frac{\rho}{2}\|x\|_{V^*}^2$, Theorem 5.1 shows that $\sigma_{\rho,T}$ is smooth and convex with Lipschitz-continuous gradient $\nabla\sigma_{\rho,T}(z) = x_{\rho,z}$, where $x_{\rho,z}$ denotes the unique optimal solution of the problem (5.1). The Lispchitz constant of this gradient is given by $L(\sigma_{\rho,T}) = \frac{1}{\rho}$.

Remark 2. When the function $J(u)$ is strongly convex with parameter κ, we do not need to apply the first smoothing to $\phi(z)$, which is, in this case, already smooth with a Lipschitz-continuous gradient with constant $\frac{\|A\|^2}{\kappa}$.

If we denote $D_T = \max\{\frac{1}{2}\|x\|_{V^*}^2 : x \in T\}$ and $D_S = \max\{\frac{1}{2}\|u\|_U^2 : u \in S\}$, it is easy to check that $\sigma_{\rho,T}(z) \le \sigma_T(z) \le \sigma_{\rho,T}(z) + \rho D_T$ $\forall z \in V$ and $\phi_\mu(z) \le \phi(z) \le \phi_\mu(z) + \mu D_S$ for all $z \in V$. Therefore, if we define the regularized function

$$\theta_{\rho,\mu}(z) = \sigma_{\rho,T}(z) + \phi_\mu(z),$$

we have $\theta_{\rho,\mu} \in F^{1,1}_{L(\rho,\mu)}(V)$ with $L(\rho,\mu) := \frac{1}{\rho} + \frac{\|\mathcal{A}\|^2}{\mu}$ and

$$\theta_{\rho,\mu}(z) \le \theta(z) \le \theta_{\rho,\mu}(z) + \mu D_S + \rho D_T \quad \forall z \in V. \tag{5.4}$$

Applying a fast gradient method to the function $\theta_{\rho,\mu}$ will generate a point $z_\epsilon \in V$ such that $\theta(z_\epsilon) - \theta^* \le \epsilon$ in $O\left(\frac{1}{\epsilon}\right)$ iterations (see [20]). However, our aim is not only to solve the dual problem efficiently but also to generate a nearly optimal and nearly feasible solution for the primal problem. We will see that a single smoothing is not enough in order to achieve this goal.

Let us first show how it is possible to reconstruct a good primal solution from a dual iterate. Let $z \in V$, then we have

$$\theta_{\rho,\mu}(z) = \sigma_{\rho,T}(z) + \phi_\mu(z) = \langle x_{\rho,z}, z\rangle - \frac{\rho}{2}\|x_{\rho,z}\|_{V^*}^2 - J(u_{\mu,z}) - \langle \mathcal{A}u_{\mu,z}, z\rangle - \frac{\mu}{2}\|u_{\mu,z}\|_U^2 .$$

Let us find the conditions that z must satisfy in order to guarantee that $u_{\mu,z}$ is nearly optimal and nearly feasible for the primal problem. We have

$$\begin{aligned} J(u_{\mu,z}) - D^* &= \langle x_{\rho,z}, z\rangle - \frac{\rho}{2}\|x_{\rho,z}\|_{V^*}^2 - \theta_{\rho,\mu}(z) - \langle \mathcal{A}u_{\mu,z}, z\rangle - \frac{\mu}{2}\|u_{\mu,z}\|_U^2 + \theta^* \\ &= \langle \nabla\theta_{\rho,\mu}(z), z\rangle + (\theta^* - \theta_{\rho,\mu}(z)) - \frac{\rho}{2}\|x_{\rho,z}\|_{V^*}^2 - \frac{\mu}{2}\|u_{\mu,z}\|_U^2 . \end{aligned}$$

Therefore, since $J(u_{\mu,z}) \ge P^* \ge D^*$, we have $J(u_{\mu,z}) - D^* \ge J(u_{\mu,z}) - P^* \ge 0$ and

$$J(u_{\mu,z}) - P^* \le |\langle \nabla\theta_{\rho,\mu}(z), z\rangle| + |\theta_{\rho,\mu}(z) - \theta^*| + \rho D_T + \mu D_S,$$

and as we also have that $|\theta_{\rho,\mu}(z) - \theta^*| \le |\theta(z) - \theta^*| + \mu D_S + \rho D_T$, we conclude that

$$J(u_{\mu,z}) \le P^* + |\langle \nabla\theta_{\rho,\mu}(z), z\rangle| + (\theta(z) - \theta^*) + 2\rho D_T + 2\mu D_S.$$

If we apply the fast gradient method to the function $\theta_{\rho,\mu}$ with constants ρ and μ chosen to be of order $O(\frac{1}{k})$, we already know from [20] that the kth iterate z_k generated by this algorithm satisfies $\theta(z_k) - \theta^* \le O(\frac{1}{k})$. However, unfortunately, the norm of the gradient of the smoothed function $\|\nabla\theta_{\rho,\mu}(z_k)\|_V$ does not decrease at the same rate, which negatively affects the above estimate of the rate of convergence of $J(u_{\mu,z})$. Indeed, as $\theta_{\rho,\mu} \in F^{1,1}_{L(\rho,\mu)}(V)$, we have (see Theorem 2.1.5 in [19])

$$\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*}^2 \le 2L(\rho,\mu)(\theta_{\rho,\mu}(z_k) - \theta^*_{\rho,\mu}).$$

As the fast gradient method is applied to the function $\theta_{\rho,\mu}$, we also have ([20])

$$\theta_{\rho,\mu}(z_k) - \theta^*_{\rho,\mu} \le \frac{4L(\rho,\mu)\|z_0 - z_S^*\|}{(k+1)(k+2)},$$

where z_S^* denotes any optimal solution of the smoothed dual problem $\min_{z\in V}\theta_{\rho,\mu}(z)$, and we see that choosing $L(\rho,\mu)$ of order $O(k)$ guarantees convergence of the smoothed objective of order $O(\frac{1}{k})$. Therefore,

$$\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \le \frac{2\sqrt{2}L(\rho,\mu)\sqrt{\|z_0 - z_S^*\|}}{\sqrt{(k+1)(k+2)}},$$

and due to the fact that $L(\rho,\mu)$ is of order $O(k)$, we cannot guarantee that the norm of the gradient $\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*}$ is decreasing with respect to k.

In principle, this can be remedied with a minor modification of the scheme. It is indeed possible to obtain in $2k$ iterations a point $\tilde{z}$ such that $\|\nabla\theta_{\rho,\mu}(\tilde{z})\|_{V^*}$ is of order $O(\frac{1}{\sqrt{k}})$. Indeed, after k steps of the fast gradient method have been computed, we can apply k additional steps of the classical gradient method with constant stepsize $\frac{1}{L(\rho,\mu)}$, i.e.,

$$z_{k+i} = z_{k+i-1} - \frac{1}{L(\rho,\mu)}\nabla\theta_{\rho,\mu}(z_{k+i-1}), \quad i = 1,\ldots,k,$$

We then have (see Theorem 2.1.14 in [19])

$$\frac{1}{2L(\rho,\mu)}\left\|\nabla\theta_{\rho,\mu}(z_{k+i-1})\right\|_{V^*}^2 \le \theta_{\rho,\mu}(z_{k+i-1}) - \theta_{\rho,\mu}(z_{k+i}) \quad \forall i = 1,\dots,k,$$

and summing these inequalities gives

$$\begin{aligned}\sum_{i=1}^{k}\frac{1}{2L(\rho,\mu)}\left\|\nabla\theta_{\rho,\mu}(z_{k+i})\right\|_{V^*}^2 \le \theta_{\rho,\mu}(z_k) - \theta_{\rho,\mu}(z_{2k}) &\le \theta_{\rho,\mu}(z_k) - \theta^*_{\rho,\mu}\\ &\le \frac{4L(\rho,\mu)\left\|z_0 - z_S^*\right\|_V^2}{(k+1)(k+2)}.\end{aligned}$$

If we denote by $\tilde z$ the iterate with the smallest norm of the gradient, we conclude that

$$\left\|\nabla\theta_{\rho,\mu}(\tilde z)\right\|_{V^*}^2 = \min_{i=1,\dots,k}\left\|\nabla\theta_{\rho,\mu}(z_{k+i})\right\|_{V^*}^2 \le \frac{8L^2(\rho,\mu)\left\|z_0 - z_S^*\right\|_V^2}{k(k+1)(k+2)}.$$

In conclusion, after $2k$ iterates, we are capable of mixing fast and classical gradient methods to obtain a point $\tilde z$ such that $\theta_{\rho,\mu}(\tilde z) - \theta^*_{\rho,\mu} \le O(\frac{1}{k})$ and $\left\|\nabla\theta_{\rho,\mu}(\tilde z)\right\|_{V^*} = O(\frac{1}{\sqrt{k}})$. However, this convergence is very slow, as it implies we need at least $k = O(\frac{1}{\epsilon^2})$ iterations in order to have a primal solution $u_{\mu,z_k} \in S$ with accuracy ϵ (i.e., $J(u_{\mu,z_k}) \le P^* + \epsilon$). This is not better than the result of the classical subgradient approach.

Furthermore, we must also examine the feasibility of the reconstructed primal iterate $u_{\mu,\tilde z}$. The norm of $\left\|\nabla\theta_{\rho,\mu}(\tilde z)\right\|_{V^*}$ also provides an upper bound for the non-admissibility measure of $u_{\mu,\tilde z}$. Indeed, denoting $d(\cdot, T)$ the distance to set T, we have

$$d(\mathcal{A}u_{\mu,\tilde z}, T) \le \left\|\mathcal{A}u_{\mu,\tilde z} - x_{\rho,\tilde z}\right\|_V = \left\|\nabla\theta_{\rho,\mu}(\tilde z)\right\|_V.$$

In conclusion, we have shown why requiring a good convergence rate for $\theta(z_k) - \theta^*$ alone is not sufficient to obtain a nearly feasible and optimal solution for the primal problem. The same good rate must also be ensured for the norm of the gradient $\left\|\nabla\theta_{\rho,\mu}(z_k)\right\|_V$. We now show how applying a second smoothing to the dual objective function, making it also strongly convex, will achieve this goal.

5.2. Second smoothing. In order to obtain a strongly convex dual objective function, we simply add the strongly convex function $\frac{\kappa}{2}\left\|z\right\|_V^2$ to the function $\theta_{\rho,\mu}$. This gives us a new dual objective function:

$$\theta_{\rho,\mu,\kappa}(z) = \sigma_{\rho,T}(z) + \phi_\mu(z) + \frac{\kappa}{2}\left\|z\right\|_V^2,$$

which is strongly convex with parameter κ. If we denote by $B = B^* : V \to V^*$ (with $B \succ 0$) the duality map between V and its dual space, i.e., $\langle Bz, \overline{z}\rangle = (z|\overline{z})_V \ \forall z, \overline{z} \in V$, we have $\nabla\theta_{\rho,\mu,\kappa}(z) = x_{\rho,z} - \mathcal{A}u_{\mu,z} + \kappa Bz$. This gradient is Lipschitz-continuous with constant $L(\rho,\mu,\kappa) := \frac{1}{\rho} + \frac{\|\mathcal{A}\|^2}{\mu} + \kappa$, hence this function belongs to $S^{1,1}_{\kappa,L(\rho,\mu,\kappa)}(V)$. Denote by $\theta^*_{\rho,\mu,\kappa}$ the optimal value of problem $\min_{z\in S}\theta_{\rho,\mu,\kappa}(z)$. Applying a fast gradient method for the class $S^{1,1}_{\kappa,L(\rho,\mu),\kappa}$ to function $\theta_{\rho,\mu,\kappa}$, we generate a sequence z_k satisfying

$$\theta_{\rho,\mu,\kappa}(z_k) - \theta^*_{\rho,\mu,\kappa} \overset{(2.3)}{\le} \exp\left(-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}\right)2(\theta_{\rho,\mu,\kappa}(z_0) - \theta^*_{\rho,\mu,\kappa})$$

and

$$\|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*}^2 \overset{(2.4)}{\leq} 4L(\rho,\mu,\kappa)(\theta_{\rho,\mu,\kappa}(z_0) - \theta^*_{\rho,\mu,\kappa}) \exp\left(-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}\right);$$

i.e.,

$$\|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} \leq 2\sqrt{L(\rho,\mu,\kappa)}\sqrt{\theta_{\rho,\mu,\kappa}(z_0) - \theta^*_{\rho,\mu,\kappa}} \exp\left(-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}\right)$$

and we have the same rate of convergence for $\|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*}$ as for $\theta_{\rho,\mu,\kappa}(z_k)-\theta^*_{\rho,\mu,\kappa}$. This property is crucial in order to obtain a nearly feasible and optimal primal solution in $O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)$ iterations (instead of $O\left(\frac{1}{\epsilon^2}\right)$ with a simple smoothing).

6. Strong duality and norm of dual optimal solutions. Before we apply the fast gradient method to the doubly smoothed dual function, we study in this section under which condition strong duality, i.e., $P^* = D^*$, holds and how it is possible to bound the size of the optimal solution set of the dual problem (3.3). Such a bound will play a role in the convergence analysis of our scheme, as we will see in the following section.

THEOREM 6.1. *If there exists $r > 0$ such that*

$$B(0,r) \subset Q := \{x - \mathcal{A}u : u \in S, x \in T\} \subset V^*, \tag{6.1}$$

then

- *there is no duality gap, i.e., $P^* = D^*$, and*
- *the optimal solution set of the dual problem (3.3) is a nonempty, bounded, closed, and convex set in V.*

Furthermore, if we define $\Delta(J') := \max_{u,v\in S} J'(v, u-v)$, where $J'(v,u-v)$ is the directional derivative of J at point v in direction $u-v$, we have the following upper-bound for the norm of the dual optimal solutions:

$$\|z^*\|_V \leq \frac{\Delta(J')}{r}.$$

Proof. Applying Theorem 2.165 in [4] to the primal problem $\min_{u\in U}\{f(u) = J(u) + I_S(u) : G(u) = Au \in T\}$ (where I_S denotes the indicator function of S), we conclude, using our assumptions on J, A, S, and T and the regularity condition (6.1), that there is no duality gap between this problem and its Lagrangian dual (3.3). Furthermore, as the primal optimal value P^* is assumed to be finite, we have that the optimal solution set of the dual problem is a nonempty, bounded, closed, and convex set in V.

It remains to obtain the bound $\|z^*\|_V \leq \frac{\Delta(J')}{r}$. As the subproblems $\sup_{x\in T}\langle x, z\rangle$ and $\sup_{u\in S}\{-J(u) - \langle \mathcal{A}u, z\rangle\}$ are solvable for all $z \in V$,

$$\partial\theta(z) = \{x_z - \mathcal{A}u_z \in V^* : x_z \in T, \langle x_z, z\rangle = \sigma_T(z), u_z \in S, -J(u_z) - \langle \mathcal{A}u_z, z\rangle = \phi(z)\}$$

is nonempty for all $z \in V$. Let $g_z = x_z - \mathcal{A}u_z$ be any element in $\partial\theta(z)$. By the optimality condition of the problems defining $\sigma_T(z)$ and $\phi(z)$, we write

$$\langle x - x_z, z\rangle \leq 0 \quad \forall x \in T \qquad -J'(u_z, u-u_z) - \langle \mathcal{A}(u-u_z), z\rangle \leq 0 \quad \forall u \in S.$$

This can be rewritten as

$$\langle x_z, z\rangle \geq \langle x, z\rangle \quad \forall x \in T \qquad -\langle \mathcal{A}u_z, z\rangle \geq -J'(u_z, u-u_z) + \langle -\mathcal{A}u, z\rangle \quad \forall u \in S,$$

and therefore

$$\langle g_z, z\rangle = \langle x_z - \mathcal{A}u_z, z\rangle \geq \langle x, z\rangle - J'(u_z, u-u_z) + \langle -\mathcal{A}u, z\rangle = \langle x - \mathcal{A}u, z\rangle - J'(u_z, u-u_z).$$

We obtain

$$\langle x - Au, z\rangle \leq J'(u_z, u-u_z) + \langle g_z, z\rangle \quad \forall u \in S, \quad \forall x \in T,$$

and therefore

$$\max_{u\in S, x\in T} \langle x - \mathcal{A}u, z\rangle \leq \max_{u\in S} J'(u_z, u-u_z) + \langle g_z, z\rangle \leq \max_{u,v\in S} J'(v, u-v) + \langle g_z, z\rangle$$

As $B(0,r) \subset \{x - \mathcal{A}u : u \in S, x \in T\} \subset V^*$, we have $\max_{u\in S, x\in T}\langle x - \mathcal{A}u, z\rangle \geq r\,\|z\|_V$, and therefore

$$r\,\|z\|_V \leq \max_{u,v\in S} J'(v, u-v) + \langle g_z, z\rangle \quad \forall z \in V. \tag{6.2}$$

Consider now an optimal solution z^* of the dual problem. By the optimality condition of this problem, we have $0 \in \partial\theta(z^*)$, and therefore $\|z^*\|_V \leq \frac{\Delta(J')}{r}$. ∎

Remark 3. As J is convex, we have $J(u) \geq J(v) + J'(v, u-v)$ $\forall u, v \in S$ and

$$\max_{u,v\in S} J(u) - J(v) = \max_{u\in S} J(u) - \min_{v\in S} J(v) \geq \max_{u,v\in S} J'(v, u-v).$$

The condition $\max_{u,v\in S} J'(v, u-v) < +\infty$ is therefore satisfied as soon as J has bounded variation on S, i.e., $\max_{u\in S} J(u) - \min_{v\in S} J(v) < +\infty$. However, our condition is strictly weaker, as the following example shows. Consider the function

$$J(u) = \int_0^1 \ln(1 - u^2(t))dt$$

defined on $S = \{u \in L^2([0,1]) : -1 \leq u(t) \leq 1 \quad \text{a.e} \quad \text{in} \quad [0,1]\}$. Clearly $\max_{u\in S} J(u) - \min_{v\in S} J(v) = +\infty$, and this function can be seen as a barrier function for S. However, we have

$$J'(v, u-v) = \int_0^1 \frac{2(u(t) - v(t))v(t)}{1 - v^2(t)} dt$$

and $\Delta(J') = \max_{u,v\in S'} J'(v, u-v) \leq 2$.

Remark 4. If the primal problem is feasible, it is clear that there exist $\overline{u} \in S$ and $\overline{x} \in T$ such that $\mathcal{A}\overline{u} = \overline{x}$, i.e., $0 \in Q$. In order to have $B(0,r) \subset Q$ with $r > 0$, one of the following extra assumptions is enough:

- **The set T has a nonempty interior and there exists $\overline{u} \in S$ such that $\mathcal{A}\overline{\mathbf{u}} = \overline{\mathbf{x}} \in \mathbf{int\ T}$ (generalized Slater condition).**
  As $\overline{x} \in \text{int } T$, there exists $r > 0$ such that: $\overline{x} + B(0,r) \subset T$, and therefore $B(0,r) = \overline{x} - \mathcal{A}\overline{u} + B(0,r) \subset Q$.
- **The set S has a nonempty interior, $\mathcal{A} : \mathbf{U} \to \mathbf{V}^*$ is surjective and there exists $\overline{\mathbf{u}} \in \mathbf{int\ S}$ such that $\mathcal{A}\overline{\mathbf{u}} = \overline{\mathbf{x}} \in \mathbf{T}$.**
  Indeed in this case, as $\overline{u} \in \text{int } S$, there exists $\tilde{r} > 0$ such that $B(\overline{u}, \tilde{r}) \subset S$. By the Banach–Schauder theorem, the image of any open subset of U by $\mathcal{A}$ is an open subset in V^*. Therefore, there exists $r > 0$ such that $\mathcal{A}\overline{u} + B(0,r) \subset \mathcal{A}(B(\overline{u}, \tilde{r})) \subset \mathcal{A}(S)$. We conclude that $\overline{x} - \mathcal{A}\overline{u} + B(0,r) = B(0,r) \subset Q$.

7. Solving the primal-dual problem in $O(\frac{1}{\epsilon}\ln(\frac{1}{\epsilon}))$ iterations. Denote by z^*_{DS} the unique optimal solution of problem

$$\min_{z\in V} \theta_{\rho,\mu,\kappa}(z), \tag{7.1}$$

and by z^* one of the optimal solutions of the dual problem (3.3). We assume that the upper bound

$$\|z^*\|_V \le D_D \tag{7.2}$$

is available. As we have just shown, this can be ensured under very natural assumptions on S, T, and J using Theorem 6.1.

If we apply method (2.2) to the doubly smoothed dual problem with starting point $z_0 = 0$, we obtain a sequence $\{z_k\}$ verifying

$$\begin{array}{rcll} \theta_{\rho,\mu,\kappa}(z_k) - \theta_{\rho,\mu,\kappa}(z^*_{DS}) & \le & 2 & (\theta_{\rho,\mu,\kappa}(0) - \theta_{\rho,\mu,\kappa}(z^*_{DS})) \quad e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}, \\ \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|^2_{V^*} & \le & 4L(\rho,\mu,\kappa) & (\theta_{\rho,\mu,\kappa}(0) - \theta_{\rho,\mu,\kappa}(z^*_{DS})) \quad e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}, \\ \|z_k - z^*_{DS}\|^2_{V^*} & \le & \min\{\|z^*_{DS}\|^2_{V^*}, \frac{4}{\kappa} & (\theta_{\rho,\mu,\kappa}(0) - \theta_{\rho,\mu,\kappa}(z^*_{DS})) \quad e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}\}. \end{array} \tag{7.3}$$

7.1. Convergence of $\theta(\mathbf{z_k})$ to θ^*. Since $\theta_{\rho,\mu,\kappa}(0) = \theta_{\rho,\mu}(0)$ and $\theta_{\rho,\mu,\kappa}(z^*_{DS}) = \theta_{\rho,\mu}(z^*_{DS}) + \frac{\kappa}{2}\|z^*_{DS}\|^2_V$, we have

$$\begin{array}{rcl} \frac{\kappa}{2}\|z^*_{DS}\|^2_V & \le & \theta_{\rho,\mu,\kappa}(0) - \theta_{\rho,\mu,\kappa}(z^*_{DS}) \; = \; \theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}) - \frac{\kappa}{2}\|z^*_{DS}\|^2_V, \\ \|z_k - z^*_{DS}\|^2_V & \overset{(2.3)}{\le} & \frac{4}{\kappa}(\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}))e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}. \end{array} \tag{7.4}$$

Note that

$$\theta_{\rho,\mu}(z_k) - \theta_{\rho,\mu}(z^*_{DS}) \overset{(2.3)}{\le} 2(\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}))e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}} + \frac{\kappa}{2}(\|z^*_{DS}\|^2_V - \|z_k\|^2_V).$$

On the other hand,

$$\begin{array}{rcl} \|z^*_{DS}\|^2_V - \|z_k\|^2_V & \le & \|z^*_{DS} - z_k\|_V(\|z^*_{DS}\|_V + \|z_k\|_V) \\ & \le & \|z^*_{DS} - z_k\|_V(2\|z^*_{DS}\|_V + \|z_k - z^*_{DS}\|_V) \\ & \overset{(2.5)}{\le} & 3\|z^*_{DS} - z_k\|_V \cdot \|z^*_{DS}\|_V \\ & \overset{(7.4)}{\le} & 3 \cdot \|z^*_{DS}\|_V \sqrt{\frac{4}{\kappa}(\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}))}e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}} \\ & \overset{(7.4)}{\le} & \frac{6}{\kappa}(\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}))e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}, \end{array}$$

and therefore

$$\theta_{\rho,\mu}(z_k) - \theta_{\rho,\mu}(z^*_{DS}) \quad \le \quad 5(\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}))e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}.$$

We also have $\theta_{\rho,\mu}(0) \le \theta(0)$ and

$$\theta_{\rho,\mu}(z^*_{DS}) \ge \theta(z^*_{DS}) - \rho D_T - \mu D_S \;\ge\; \theta(z^*) - \rho D_T - \mu D_S.$$

Therefore,

$$\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}) \quad \le \quad \theta(0) - \theta(z^*) + \rho D_T + \mu D_S. \tag{7.5}$$

Finally, since $\theta_{\rho,\mu}(z^*_{DS}) + \frac{\kappa}{2}\|z^*_{DS}\|^2_V \le \theta_{\rho,\mu}(z^*) + \frac{\kappa}{2}\|z^*\|^2_V$, we have

$$\theta_{\rho,\mu}(z^*_{DS}) \le \theta_{\rho,\mu}(z^*) + \frac{\kappa}{2}\|z^*\|^2_V \overset{(5.4)}{\le} \theta(z^*) + \frac{\kappa}{2}\|z^*\|^2_V,$$

and therefore

$$\theta_{\rho,\mu}(z_k) - \theta_{\rho,\mu}(z^*_{DS}) \overset{(5.4)}{\ge} \theta(z_k) - \mu D_S - \rho D_T - \theta(z^*) - \tfrac{\kappa}{2}\|z^*\|^2_V.$$

In conclusion, we have

$$\theta(z_k) - \theta(z^*) \le \mu D_S + \rho D_T + \tfrac{\kappa}{2} D_D^2 + 5\left(\theta(0) - \theta(z^*) + \rho D_T + \mu D_S\right) e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}. \tag{7.6}$$

Now it is clear how to choose the smoothing parameters. Let us fix some $\epsilon > 0$. In the upper bound for the residual $\theta(z_k) - \theta(z^*)$, we have four terms. In order to ensure accuracy $\theta(z_k) - \theta(z^*) \le \epsilon$, we force all of these terms to be less than or equal to $\frac{\epsilon}{4}$. This leads to the following values:

$$\mu = \mu(\epsilon) = \tfrac{\epsilon}{4D_S}, \quad \rho = \rho(\epsilon) = \tfrac{\epsilon}{4D_T}, \quad \kappa = \kappa(\epsilon) = \tfrac{\epsilon}{2D_D^2}. \tag{7.7}$$

Under this choice we get

$$\theta(z_k) - \theta(z^*) \le \frac{3\epsilon}{4} + 5\left(\theta(0) - \theta(z^*) + \frac{\epsilon}{2}\right) e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}. \tag{7.8}$$

The last term in the estimate (7.8) defines the number of iterations needed for reaching the accuracy ϵ. Clearly, we ensure $5\left(\theta(0) - \theta(z^*) + \frac{\epsilon}{2}\right) e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}} \le \frac{\epsilon}{4}$ by taking

$$k \ge 2\sqrt{\frac{L(\rho,\mu,\kappa)}{\kappa}} \ln \frac{20\left(\theta(0) - \theta(z^*) + \frac{\epsilon}{2}\right)}{\epsilon}. \tag{7.9}$$

It remains to note that

$$\frac{L(\rho,\mu,\kappa)}{\kappa} = 1 + \frac{1}{\rho\kappa} + \frac{1}{\mu\kappa}\|\mathcal{A}\|^2 \overset{(7.7)}{=} 1 + \frac{8}{\epsilon^2}\left[D_T + D_S\|\mathcal{A}\|^2\right] D_D^2. \tag{7.10}$$

Thus, we need at most $k = O(\frac{1}{\epsilon}\ln\frac{1}{\epsilon})$ iterations.

7.2. Convergence of $\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*}$. In our approach, we want to be able to reconstruct a nearly optimal and feasible primal solution efficiently. In section 4.1, we have seen that the accuracy of this primal solution depends not only on the rate of convergence for the dual objective function, but also on the rate of convergence of the norm of its gradient. We now provide an upper bound on the number of iterations needed to reduce this norm below a certain level. We start with

$$\begin{aligned}\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} &\le \|\nabla\theta_{\rho,\mu,\kappa}(z_k) - \kappa B z_k\|_{V^*} \le \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} + \kappa\|Bz_k\|_{V^*} \\ &= \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} + \kappa\|z_k\|_V \overset{(7.3)}{\le} \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} + 2\kappa\left\|z^*_{DS}\right\|_V.\end{aligned}$$

Note that

$$\begin{aligned}\tfrac{1}{4L(\rho,\mu,\kappa)}\left\|\nabla\theta_{\rho,\mu,\kappa}(z_k)\right\|^2_{V^*} &\overset{(7.3),(7.4)}{\le} (\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}))e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}} \\ &\overset{(7.5)}{\le} (\theta(0) - \theta(z^*) + \mu D_S + \rho D_T)e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}} \\ &\overset{(7.7)}{=} (\theta(0) - \theta(z^*) + \tfrac{\epsilon}{2})e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}.\end{aligned}$$

At the same time,

$$\begin{aligned}\theta(z^*) + \tfrac{\kappa}{2}\|z^*\|_V^2 &\overset{(5.4)}{\geq} \theta_{\rho,\mu}(z^*) + \tfrac{\kappa}{2}\|z^*\|_V^2 \ \geq\ \theta_{\rho,\mu}(z^*_{DS}) + \tfrac{\kappa}{2}\|z^*_{DS}\|_V^2\\ &\overset{(5.4)}{\geq} \theta(z^*_{DS}) - \mu D_S - \rho D_T + \tfrac{\kappa}{2}\|z^*_{DS}\|_V^2\\ &\geq\ \theta(z^*) - \mu D_S - \rho D_T + \tfrac{\kappa}{2}\|z^*_{DS}\|_2^2.\end{aligned}$$

Hence,

$$\|z^*_{DS}\|_V \leq \sqrt{\|z^*\|_2^2 + \frac{2\mu}{\kappa}D_S + \frac{2\rho}{\kappa}D_T} \overset{(7.7)}{\leq} \kappa^{-1/2}\sqrt{\frac{3\epsilon}{2}} \overset{(7.7)}{=} \sqrt{3}D_D, \tag{7.11}$$

and we obtain

$$\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \leq \sqrt{4L(\rho,\mu,\kappa)\left(\theta(0) - \theta(z^*) + \frac{\epsilon}{2}\right)}e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}} + 2\sqrt{3}\kappa D_D.$$

Taking into account (7.7), we can see that in $k(\epsilon) = O(\frac{1}{\epsilon}\ln\frac{1}{\epsilon})$ iterations, we can ensure

$$\theta(z_k) - \theta(z^*) \leq \epsilon, \quad \|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \leq \tfrac{2\epsilon}{D_D}. \tag{7.12}$$

7.3. Constructing an approximate primal solution. In this section, given a target accuracy $\epsilon > 0$, we will see how to obtain from the dual iterate $z_{k(\epsilon)}$ an approximate primal solution $\widehat{u}_{k(\epsilon)} \in S$ such that

$$\left|J(\widehat{u}_{k(\epsilon)}) - D^*\right| \quad\leq\quad 2\left(1 + 2\sqrt{3}\right)\cdot\epsilon, \tag{7.13}$$

$$d(\mathcal{A}\widehat{u}_{k(\epsilon)}, T) \leq \frac{2\epsilon}{D_D}. \tag{7.14}$$

Since $D^* \leq P^*$, inequality (7.13) implies $J(\widehat{u}_{k(\epsilon)}) \leq P^* + 2\left(1 + 2\sqrt{3}\right)\cdot\epsilon$, and $\hat{u}_{k(\epsilon)}$ satisfying (7.13), (7.14) can be seen as a nearly optimal and feasible primal solution with accuracy proportional to ϵ.

Consider $\widehat{u}_{k(\epsilon)} = u_{\mu(\epsilon),z_{k(\epsilon)}}$, the unique optimal solution of the optimization problem defining $\phi_{\mu(\epsilon)}(z_{k(\epsilon)})$. We have

$$\begin{aligned}\theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)}) &= \sigma_{\rho(\epsilon),Q}(z_{k(\epsilon)}) + \phi_\mu(z_{k(\epsilon)})\\ &= \langle x_{\rho(\epsilon),z_{k(\epsilon)}}, z_{k(\epsilon)}\rangle - \tfrac{\rho(\epsilon)}{2}\|x_{\rho(\epsilon),z_{k(\epsilon)}}\|_{V^*}^2 - J(\widehat{u}_{k(\epsilon)})\\ &\quad -\langle\mathcal{A}\widehat{u}_{k(\epsilon)}, z_{k(\epsilon)}\rangle - \tfrac{\mu(\epsilon)}{2}\|\widehat{u}_{k(\epsilon)}\|_U^2.\end{aligned}$$

Therefore,

$$\begin{aligned}J(\widehat{u}_{k(\epsilon)}) - D^* &= \langle x_{\rho(\epsilon),z_{k(\epsilon)}} - \mathcal{A}\widehat{u}_{k(\epsilon)}, z_{k(\epsilon)}\rangle - \tfrac{\rho(\epsilon)}{2}\|x_{\rho(\epsilon),z_{k(\epsilon)}}\|_{V^*}^2\\ &\quad -\tfrac{\mu(\epsilon)}{2}\left\|\widehat{u}_{k(\epsilon)}\right\|_U^2 - \theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)}) + \theta(z^*).\end{aligned} \tag{7.15}$$

Since $\theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)}) - \theta(z^*) \leq \theta(z_{k(\epsilon)}) - \theta(z^*) \leq \epsilon$, and

$$\begin{aligned}\theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)}) - \theta(z^*) &\overset{(5.4)}{\geq} \theta(z_{k(\epsilon)}) - \mu(\epsilon)D_S - \rho(\epsilon)D_T - \theta(z^*)\\ &\overset{(7.7)}{=} \theta(z_{k(\epsilon)}) - \theta(z^*) - \tfrac{1}{2}\epsilon \ \geq\ -\tfrac{1}{2}\epsilon,\end{aligned}$$

we have $|\theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)}) - \theta(z^*)| \le \epsilon$. Therefore,

$$\begin{aligned}\left|J(\widehat{u}_{k(\epsilon)}) - D^*\right| &\le \left\|x_{\rho(\epsilon),z_{k(\epsilon)}} - \mathcal{A}\widehat{u}_{k(\epsilon)}\right\|_{V^*} \left\|z_{k(\epsilon)}\right\|_V + \rho(\epsilon)\hat{D}_T + \mu(\epsilon)D_S + \epsilon \\ &\overset{(7.7)}{\le} \left\|\nabla\theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)})\right\|_{V^*} \left\|z_{k(\epsilon)}\right\|_V + 2\epsilon \overset{(7.12)}{\le} \tfrac{2\epsilon}{D_D}\left\|z_{k(\epsilon)}\right\|_V + 2\epsilon,\end{aligned}$$

establishing (7.13). On the other hand,

$$\left\|z_{k(\epsilon)}\right\|_V \le \left\|z_{k(\epsilon)} - z^*_{DS}\right\|_V + \|z^*_{DS}\|_V \overset{(7.3)}{\le} 2\,\|z^*_{DS}\|_V \overset{(7.11)}{\le} 2\sqrt{3}D_D$$

and we obtain $\left|J(\widehat{u}_{k(\epsilon)}) - D^*\right| \le 2\left(1+2\sqrt{3}\right)\cdot\epsilon$. Finally, we have $\widehat{u}_{k(\epsilon)} \in S$ and

$$\left\|\mathcal{A}\widehat{u}_{k(\epsilon)} - x_{\rho(\epsilon),z_{k(\epsilon)}}\right\|^2_{V^*} = \|\nabla\theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)})\|^2_{V^*} \overset{(7.12)}{\le} \left(\tfrac{2\epsilon}{D_D}\right)^2,$$

where $x_{\rho(\epsilon),z_{k(\epsilon)}} \in T$, which proves (7.14).

8. Practical implementation of the double smoothing technique. The double smoothing technique presented in the previous section relies on a choice of the smoothing parameter κ that requires knowledge of the size of the dual optimal set, or at least an upper-bound D_D on this quantity (given, for example, by Theorem 6.1).

Even when we have no initial estimate of the size of $\|z^*\|_V$, we can still apply a simple modification of the double smoothing technique. The procedure we propose is based on a initial estimate $\overline{D}_D$ of $\|z^*\|_V$ (we do not require $\|z^*\|_V \le \overline{D}_D$). It consists of successively applying the fast gradient method to a sequence of doubly smoothed dual objective functions $\theta_{\rho,\mu,\kappa}(\cdot)$ with decreasing smoothing parameter κ, restarting from scratch at each application, until condition $\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \le \varepsilon$ is satisfied. This procedure will ensure we obtain a primal iterate $u \in S$ that is both nearly feasible for the linear constraints, i.e., $d(\mathcal{A}u, T) \le O(\varepsilon)$, and nearly optimal for the primal objective function, i.e., $J(u) - P^* \le O(\max(\|z^*\|_V, \overline{D}_D)\varepsilon)$. Moreover, when compared to the ideal situation where the norm of the dual optimal solution is known, the total number of iterations required by this procedure only grows by a small logarithmic multiplicative factor of order $O(\max(1, \log_2(\frac{\|z^*\|_V}{\overline{D}_D})))$.

Let $\varepsilon > 0$. The proposed global procedure goes as follows.

Initialization: Let $\overline{D}_D$ be an initial estimate for $\|z^*\|_V$. Choose $\kappa(1) = \frac{\varepsilon}{8\overline{D}_D}$ and $j = 1$.

jth attempt; Apply (from scratch) the fast gradient method to the doubly smoothed dual function $\theta_{\rho,\mu,\kappa}(.)$ with $\mu = \frac{\varepsilon^2}{128 D_S \kappa(j)}$ and $\rho = \frac{\varepsilon^2}{128 D_T \kappa(j)}$ until the following stopping criterion is met:

$$\|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} \le \frac{\varepsilon}{4}. \tag{8.1}$$

If this last iterate satisfie

$$\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \le \varepsilon, \tag{8.2}$$

terminate and output $u_k = \arg\min_{u\in U}\{J(u) + \langle\mathcal{A}u, z_k\rangle + \frac{\mu}{2}\|u\|^2_U\}$.
Else let $\kappa(j+1) = \frac{1}{2}\kappa(j)$ and start the $(j+1)$th attempt.

First, we note that since the stopping criterion (8.1) checked at the jth attempt is the norm of the doubly regularized function being minimized, it is guaranteed to be met after at most $k = \sqrt{1+\frac{128D_T}{\varepsilon^2}+\frac{128D_S}{\varepsilon^2}}\log(\frac{16\kappa(j)(1+\frac{128D_T}{\varepsilon^2}+\frac{128D_S}{\varepsilon^2})(\theta(0)-\theta^*+\frac{\varepsilon^2}{64\kappa})}{\varepsilon})$ iterations; see bound (2.4).

We now need to estimate the number of attempts needed to reach the termination condition (8.2), and consider two cases:

1. When $\overline{D}_D$ is chosen smaller than the true norm of the dual optimal solution $\|z^*\|_V$, our global process stops for sure with criterion (8.2) when

$$\frac{\varepsilon}{16\|z^*\|_V} \leq \kappa(j) \leq \frac{\varepsilon}{8\|z^*\|_V}. \tag{8.3}$$

Indeed, as soon as $\kappa(j)$ satisfies this condition we have

$$\begin{aligned}
\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} &\leq \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} + \kappa\|z_k\|_V \\
&\overset{(2.5),(7.11)}{\leq} \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} + 2\kappa\sqrt{\|z^*\|_V^2 + \frac{2\mu}{\kappa}D_S + \frac{2\rho}{\kappa}D_T} \\
&\leq \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} + 2\kappa\|z^*\|_V + 2\sqrt{2}\sqrt{\kappa\mu D_S} + 2\sqrt{2}\sqrt{\kappa\rho D_T} \\
&\overset{(8.1),(8.3)}{\leq} \varepsilon.
\end{aligned}$$

As $\kappa(1) = \frac{\varepsilon}{8\overline{D}_D}$ and $\kappa(j) = \frac{1}{2}\kappa(j-1)$, the number of attempts carried out before condition (8.3) holds, satisfies $\log_2(\frac{\|z^*\|_V}{\overline{D}_D}) + 1 \leq j \leq \log_2(\frac{\|z^*\|_V}{\overline{D}_D}) + 2$.

When this global procedure stops, primal iterate u_k is nearly feasible,

$$d(\mathcal{A}u_k, T) \leq \|\mathcal{A}u_k - x_{\rho,z_k}\|_{V^*} = \|\nabla\theta_{\rho,\mu}(z_k)\|_V^* \leq \varepsilon,$$

and nearly optimal,

$$\begin{aligned}
J(u_k) - P^* &\overset{(7.15)}{\leq} |\langle\nabla\theta_{\rho,\mu}(z_k), z_k\rangle| + \theta^* - \theta_{\rho,\mu}(z_k) \\
&\leq \|\theta_{\rho,\mu}(z_k)\|_{V^*}\|z_k\|_V + \rho D_T + \mu D_S \\
&\overset{(2.5),(7.11),(8.2)}{\leq} 2\varepsilon\sqrt{\|z^*\|_V^2 + \frac{2\mu D_S}{\kappa} + \frac{2\rho D_T}{\kappa}} + \rho D_T + \mu D_S \\
&\overset{(8.3)}{\leq} \frac{25}{4}\varepsilon\|z^*\|_V.
\end{aligned}$$

2. If $\overline{D}_D$ is chosen bigger than $\|z^*\|_V$, the global procedure stops after the first attempt since $\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \leq \frac{3\varepsilon}{4} + \frac{\varepsilon\|z^*\|_V}{4\overline{D}_D} \leq \varepsilon$, and we obtain a primal iterate u_k satisfying $d(\mathcal{A}u_k, T) \leq \|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \leq \varepsilon$ and $J(u_k) - P^* \leq 2\varepsilon\sqrt{\|z^*\|_V^2 + 2\overline{D}_D^2} + \frac{\overline{D}_D\varepsilon}{8} \leq (2\sqrt{3}+\frac{1}{8})\overline{D}_D\varepsilon$.

In conclusion, when the global procedure stops, we have an approximately feasible solution with $d(\mathcal{A}u_k, T) \leq \varepsilon$ and a guarantee on the accuracy of the objective function of order $O(\max(\|z^*\|_V, \overline{D}_D)\varepsilon)$. Although this last bound cannot usually be computed in practice, since $\|z^*\|_V$ is not known, an a posteriori guarantee on the objective accuracy can be easily computed from the right-hand side of inequality $J(u_k) - P^* \leq |\langle\nabla\theta_{\rho,\mu}(z_k), z_k\rangle| + \rho D_T + \mu D_S$.

Remark 5. A potential drawback of the suggested scheme is that while it allows the user to choose a guarantee ε on the feasibility of the final iterate, it does not allow the same with the objective function accuracy. Indeed, our bound for that accuracy $O(\max(\|z^*\|_V, \overline{D}_D)\varepsilon)$ involves the unknown norm of the dual optimal solution.[1] However, this is unavoidable, as all known methods for convex optimization require at least some information of this type to guarantee absolute accuracy of the solution. For example, the standard analysis of the primal gradient method, which produces inequality $f(x_k) - f^* \leq \frac{Ld(x_0, X^*)}{2k}$, requires knowledge of the initial distance to the primal optimal solution set X^*, whose nature is similar to the size of the primal optimal solution.

9. Applications in optimal control. In this section, we will study optimal control problems (OCP) that can be written in the form (3.1) (more precisely in the form (4.1)). In particular, we consider OCP governed by a system of linear differential equations with convex objective functional, convex constraints on the state variables at finite number of *inspection moments*, and pointwise convex constraints on the control variables. In order to motivate our choice of problem class, in particular the linearity of the differential equations, we first show that OCP with a nonlinear system of differential equations are NP-hard.

Consider the following OCP with convex objective function:

$$\min_{u \in L^2([0,1], R^n)} \left\{ \|x(1)\|_4^4 + \langle c, x(1)\rangle^2 : \dot{x} = -x \cdot \langle x, u\rangle + u, \ x(0) = x_0 \right\}. \tag{9.1}$$

We assume that vector c has integer coefficients.

LEMMA 9.1. *Let $\|x_0\|_2^2 = 1$. Then, finding an approximate solution to problem* (9.1) *with absolute accuracy higher than $\hat{\epsilon} \stackrel{\mathrm{def}}{=} \frac{1}{n(1+n^{1/2}\|c\|_1)^2}$ is NP-hard.*

Proof. In view of the system of differential equations in (9.1), we have $\langle \dot{x}, x\rangle \equiv 0$. Hence, by condition of the lemma, $\|x(t)\|_2 \equiv 1$. Note that by an appropriate control u we can move the starting point x_0 to any position at the unit sphere. Hence, the problem (9.1) is equivalent to the following finite-dimensional minimization problem:

$$\text{Find } \xi_* = \min_{\|y\|_2 = 1} \left\{ \xi(y) \stackrel{\mathrm{def}}{=} \|y\|_4^4 + \langle c, y\rangle^2 \right\}. \tag{9.2}$$

Let us show that this problem is equivalent to solving the equation $\langle c, y\rangle = 0$ with Boolean variables (which is a well-known NP-hard problem).

If this equation has Boolean solution y_* with coefficients $y_*^{(i)} = \pm m^{-1/2}$, $i = 1, \dots, n$, then $\xi_* = \frac{1}{m}$. On the other hand, note that for any $y \in R^m$ with unit Euclidean norm we have $\|y\|_4^4 = \frac{1}{m} + \sum_{i=1}^m \left((y^{(i)})^2 - \frac{1}{m}\right)^2$. If we manage to find such a point y with $\xi(y) - \xi_* < \hat{\epsilon}$, then in the case $\xi(y) \geq \frac{1}{m} + \hat{\epsilon}$ we guarantee the absence of Boolean solutions. If $\xi(y) < \frac{1}{m} + \hat{\epsilon}$, then $|\langle c, y\rangle| < \hat{\epsilon}^{1/2}$ and $\max_{1 \leq i \leq m} \left| (y^{(i)})^2 - \frac{1}{m}\right| < \hat{\epsilon}^{1/2}$. In this case, we can define the Boolean vector $u^{(i)} = \frac{1}{m^{1/2}} \cdot \mathrm{sign}(y^{(i)})$, $i = 1, \dots, m$. For this vector we have

$$|\langle c, u\rangle| = |\langle c, y\rangle + \langle c, u - y\rangle| \leq |\langle c, y\rangle| + \left| \sum_{i=1}^m c^{(i)} \cdot \mathrm{sign}(y^{(i)}) \left(\tfrac{1}{n^{1/2}} - |y^{(i)}|\right)\right|$$

$$< \hat{\epsilon}^{1/2} + \|c\|_1 \max_{1 \leq i \leq m} \left| \tfrac{1}{m^{1/2}} - |y^{(i)}|\right| < \hat{\epsilon}^{1/2}(1 + m^{1/2}\|c\|_1) = \tfrac{1}{m^{1/2}}.$$

Since vector c has integer coefficients, we conclude that $\langle c, u\rangle = 0$. ∎

[1] For comparison, when D_D is known, the objective accuracy can be chosen as ϵ, and the feasibility guarantee becomes $O(\frac{\epsilon}{D_D})$, as done in the analysis of section 7.

9.1. Class of optimal control problems and reformulation. Consider the following optimal control problem:

$$\inf_u \{\int_0^1 F(t,u(t))dt : \quad \begin{aligned} &\dot{x}(t) = A(t)x(t) + B(t)u(t),\ x(0) = x_0, \\ &x(t_i) \in T_i \quad i = 1,\dots,N, \\ &u(t) \in S(t) \quad \text{a.e in } [0,1]\}, \end{aligned} \tag{9.3}$$

with variables $x(t) \in R^n$ and $u(t) \in R^m$, $t \in [0,1]$ and where sets $S(t) \subset R^m$, $t \in [0,1]$, are closed and convex with bounded graph $\overline{S} \stackrel{\text{def}}{=} \cup_{t\in[0,1]} S(t)$. We assume that function $F : [0,1] \times \overline{S} \to R$ is bounded and continuously differentiable and convex in the second argument. We measure control variables with the norm $\|u\|_2^2 = \int_0^1 \|u(t)\|_2^2 dt$. We assume that $A(t) \in \mathcal{C}([0,1], R^{n\times n})$ and $B(t) \in \mathcal{C}([0,1], R^{n\times m})$. In problem (9.3), we have a finite number of inspection moments $t_i \in (0,1]$, $i = 1,\dots,N$, and we assume that each subset $T_i \subset R^n$ is a closed, bounded convex set. Let us rewrite the problem (9.3) in terms of control u. Denote by $\Phi(t,\tau)$ the transition matrix of the system, i.e., the unique solution of the following matrix Cauchy problem:

$$\tfrac{d}{dt}\Phi(t,\tau) \quad = \quad A(t)\Phi(t,\tau), \quad t \geq \tau, \quad \Phi(\tau,\tau) = I.$$

Remark 6. When the system is time-invariant, i.e., $A(t) = A$ and $B(t) = B$ for all $t \in [0,1]$, then transition matrix Φ is the usual matrix exponent

$$\Phi(t,\tau) \quad = \quad e^{(t-\tau)A} = I + \sum_{k=1}^{\infty} \tfrac{A^k (t-\tau)^k}{k!}.$$

From classical optimal control theory (e.g., [12]), we know that the state trajectory $x(t)$, generated by the system of differential equations under the control $u(t)$, is defined by $x(t) = \Phi(t,0)x_0 + \int_0^t \Phi(t,\tau)B(\tau)u(\tau)d\tau$, $t \in [0,1]$. Therefore, the constraint $x(t_i) \in T_i$ can be expressed as $\mathcal{A}_i(u) \in \bar{T}_i$ for each $i = 1,\dots,N$ using

$$\mathcal{A}_i(u) \ \stackrel{\text{def}}{=} \ \int_0^{t_i} \Phi(t_i,\tau)B(\tau)u(\tau)d\tau \quad \text{and} \quad \overline{T}_i \stackrel{\text{def}}{=} T_i - \Phi(t_i,0)x_0, \tag{9.4}$$

where $\Phi(t_i,0)x_0$ is the value at time t_i of the unique solution of Cauchy problem

$$\dot{x}(t) = A(t)x(t), \quad x(0) = x_0.$$

Linear operator $\mathcal{A}_i : L^2([0,1], R^m) \to R^n$ can also be written as

$$\mathcal{A}_i u = \int_0^1 A_i(\tau)u(\tau)d\tau, \ \text{where } A_i(\tau) \stackrel{\text{def}}{=} \begin{cases} \Phi(t_i,\tau)B(\tau) & \text{when } \tau \in [0,t_i], \\ 0 & \text{when } \tau \in]t_i, 1]. \end{cases}$$

Defining now $J(u) = \int_0^1 F(t,u(t))dt$ and $S = \{u \in L^2([0,1], R^m) : u(t) \in S(t) \text{ a.e. in } [0,1]\}$, the optimal control problem (9.3) can be rewritten in the form (3.2), and therefore in the form (3.1), if we define the convex set $T = \bar{T}_1 \times \bar{T}_2 \times \cdots \times \bar{T}_N \subset R^{N\times n}$ and the linear operator $\mathcal{A} : L^2([0,1], R^m) \to R^{N\times n}$ such that $\mathcal{A}u : R^{N\times n} \to R$ satisfies $\langle \mathcal{A}u, z\rangle = \sum_{i=1}^N \langle \mathcal{A}_i u, z_i\rangle$ for all $u \in L^2([0,1], R^m)$ and $z = (z_1,\dots,z_N) \in R^n$. Hence, we can solve it by the double smoothing technique. This approach assumes that we are able to solve the following problems easily in a pointwise way,

$$\max_{u\in S(t)} \left\{ -F(t,u) - \sum_{i=1}^N \langle u, A_i(t)^T z^i\rangle - \tfrac{\mu}{2}\|u\|_2^2 \right\},$$

where $A_i(t)$ depends directly on the state transition matrix. However, in practice the state transition matrix $\Phi(t_i,t)$ is often not known explicitly. Instead, we can compute the function $A_i(t)^T z^i$ as a solution of some system of differential equations. Indeed, we have (see, e.g., Theorem 1.2 in [13]) $\frac{d}{dt}\Phi^T(t_i,t) = -A(t)^T\Phi^T(t_i,t)$. Therefore $\Phi(t_i,t)^T$ is the state transition matrix of the system $\dot v(t) = -A(t)^T v(t)$. Hence, $A_i(t)^T z^i = B(t)^T v(t)$, where $v(t)$ is the unique solution of Cauchy problem

$$\dot v(t) = -A(t)^T v(t), \quad v(t_i) = z^i, \quad t \in [0,t_i], \tag{9.5}$$

extended by zero for $t \in [t_i, 1]$.

Remark 7. At first glance, it seems that we are restricted to objective functionals depending only on the control $u(t)$ and not on the state variable $x(t)$. In fact, using the state transition matrix, we can also consider any convex functions depending on linear functionals of the state. Such a functional can be defined as

$$\begin{aligned} l(x) &= \textstyle\int_0^1 \langle x(t), a(t)\rangle dt \;=\; \int_0^1 \langle \int_0^t \Phi(t,\tau)B(\tau)u(\tau)d\tau, a(t)\rangle dt \\ &= \textstyle\int_0^1\int_0^t \langle \Phi(t,\tau)B(\tau)u(\tau), a(t)\rangle d\tau dt \;=\; \int_0^1\int_0^t \langle u(\tau), B(\tau)^T\Phi(t,\tau)^T a(t)\rangle d\tau dt \\ &= \textstyle\int_0^1\int_\tau^1 \langle u(\tau), B(\tau)^T\Phi(t,\tau)^T a(t)\rangle dt d\tau \;\stackrel{\text{def}}{=}\; \int_0^1 \langle u(\tau), h(\tau)\rangle d\tau, \end{aligned}$$

with $h(\tau) = \int_\tau^1 B(\tau)^T\Phi(t,\tau)^T a(t)dt$. Another possibility is as follows:

$$\begin{aligned} l(x) &= \langle x(t_i), a\rangle \;=\; \textstyle\langle \int_0^{t_i} \Phi(t_i,\tau)B(\tau)u(\tau)d\tau, a\rangle \\ &= \textstyle\int_0^{t_i} \langle \Phi(t_i,\tau)B(\tau)u(\tau), a\rangle d\tau \;\stackrel{\text{def}}{=}\; \int_0^{t_i}\langle u(\tau), h(\tau)\rangle d\tau, \end{aligned}$$

with $h(\tau) = B(\tau)^T\Phi(t_i,\tau)^T a$.

9.2. Evaluation of $\|\mathcal{A}_i\|_2$. In order to solve the primal-dual problem (3.1)–(3.3) by a double smoothing technique, we need to bound the norm $\|\mathcal{A}\|_2 \le [\sum_{i=1}^N \|\mathcal{A}_i\|_2^2]^{1/2}$. Moreover, from the estimates (7.9), (7.10), it is clear that this norm is an essential element in the global complexity bound of our problem. In this section, we first derive a closed-form representation for the norm $\|\mathcal{A}_i\|_2$ using the reachability Gramian of the dynamical system. However, this quantity is not easily computable (it needs the knowledge of the transition matrix). Moreover, its dependence in the length of time interval is not very transparent. Therefore, in the next section, we obtain some simple upper bounds for the norms $\|\mathcal{A}_i\|_2$, which can be easily computed by solving linear matrix inequalities (LMI).

Let us derive first the exact expression for $\|\mathcal{A}_i\|_2$. By definition,

$$\|\mathcal{A}_i\|_2 = \sup_{u\in L^2([0,1],R^m)} \left\{ \|\mathcal{A}_i u\|_2 : \; \|u\|_{L^2([0,1],R^m)} = 1 \right\}.$$

Since the vector $\mathcal{A}_i u$ does not depend on values of $u(t)$ for $t \in (t_i, 1]$, we can consider the restriction of $\mathcal{A}_i$ on $L^2([0,t_i],R^m) : u \to \int_0^{t_i}\Phi(t_i,\tau)B(\tau)u(\tau)d\tau$. Then

$$\|\mathcal{A}_i\|_2 = \sup_{u\in L^2([0,t_i],R^m)} \left\{ \|\mathcal{A}_i u\|_2 : \; \|u\|_{L^2([0,t_i],R^m)} = 1 \right\},$$

and the operator $\mathcal{A}_i^*$ maps $y \in R^n$ into function $B(t)^T\Phi(t_i,t)^T y \in L^2([0,t_i]$.

For all $t_i > 0$, $i = 1,\ldots,N$, define the reachability Gramians

$$W_r(0,t_i) = \int_0^{t_i} \Phi(t_i,\tau)B(\tau)B(\tau)^T\Phi(t_i,\tau)^T d\tau \;=\; \mathcal{A}_i\mathcal{A}_i^*,$$

which are symmetric positive semidefinite matrices. Recall the following definition.

DEFINITION 9.2. *The system*

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(0) = 0, \tag{9.6}$$

is called reachable *on* $[0, \hat{t}]$ *if for any* $\hat{x} \in R^n$ *there exist a control* $u(t)$ *such that* $x(\hat{t}) = \hat{x}$.

Reachability is closely related to the reachability Gramian (e.g., Corollary 2.3 in [1]).

THEOREM 9.3. *The system* (9.6) *is reachable on* $[0, \hat{t}]$ *if and only if the Gramian* $W_r(0, \hat{t})$ *is positive definite.*

Let us come back now to the definition of the norm $\|\mathcal{A}_i\|_2$. We have

$$\begin{aligned}\|\mathcal{A}_i\|_2 &= \sup_{u \in L^2([0,t_i],R^m)} \left\{ \|\mathcal{A}_i u\|_2 : \ \|u\|_{L^2([0,t_i],R^m)} = 1 \right\} \\ &= \left[\inf_{u \in L^2([0,t_i],R^m)} \left\{ \|u\|_{L^2([0,t_i],R^m)} : \ \|\mathcal{A}_i u\|_2 = 1 \right\} \right]^{-1}.\end{aligned}$$

If the system is reachable on $[0, t_i]$, then $\mathrm{Im}\ \mathcal{A}_i(L^2([0, t_i], R^m)) = R^n$, and we have

$$\begin{aligned}&\inf_{u \in L^2([0,t_i],R^m)} \left\{ \|u\|_{L^2([0,t_i],R^m)} : \ \|\mathcal{A}_i u\|_2 = 1 \right\} \\ &\quad = \inf_{\substack{x_i \in R^n, \|x_i\|_2 = 1 \\ u \in L^2([0,t_i],R^m)}} \left\{ \|u\|_{L^2([0,t_i],R^m)} : \ \mathcal{A}_i u = x_i \right\}.\end{aligned}$$

In order to solve this minimization problem, we will use the following simple result.

LEMMA 9.4. *Let* H *be a Hilbert space and let the linear operator* $A : H \to R^L$ *be nondegenerate:* $AA^* \succ 0$. *Then for any* $b \in R^L$ *and* $f \in H$, *the Euclidean projection* $\pi_b(f)$ *of* f *onto the subspace* $\mathcal{L}_b = \{g \in H : Ag = b\}$ *corresponds to* $\pi_b(f) = f + A^*(AA^*)^{-1}(b - Af)$. Therefore

$$\inf_{\substack{u \in L^2([0,t_i],R^m), \\ \mathcal{A}_i u = x_i}} \|u\|_2 = \|\mathcal{A}_i^*(\mathcal{A}_i\mathcal{A}_i^*)^{-1}x_i\|_2 \ = \langle (\mathcal{A}_i\mathcal{A}_i^*)^{-1}x_i, x_i \rangle^{1/2}.$$

Therefore

$$\begin{aligned}\inf_{u \in L^2([0,t_i],R^m)} \left\{ \|u\|_{L^2([0,t_i],R^m)} : \ \|\mathcal{A}_i u\|_2 = 1 \right\} &= \inf_{\|x_i\|_2 = 1} \langle (\mathcal{A}_i\mathcal{A}_i^*)^{-1}x_i, x_i \rangle^{1/2} \\ &= \lambda_{\min}^{1/2}((\mathcal{A}_i\mathcal{A}_i^*)^{-1}),\end{aligned}$$

and we conclude that

$$\|\mathcal{A}_i\|_2 = \lambda_{\min}^{-1/2}((\mathcal{A}_i\mathcal{A}_i^*)^{-1}) \ = \ \lambda_{\max}^{1/2}(\mathcal{A}_i\mathcal{A}_i^*),$$

where $\mathcal{A}_i\mathcal{A}_i^* = W_r(0, t_i)$ is the reachability Gramian.

9.3. Bounding the growth of norms $\|\mathcal{A}_i\|_2$ with time. In the previous section, we have shown that the norm $\|\mathcal{A}_i\|_2$ is equal to the square root of the maximal eigenvalue of the reachability Gramian on the interval $[0, t_i]$. Simple examples show that this norm can grow exponentially with t_i. However, for stable systems, the situation is much better.

In this section, we derive the bounds for the growth of the norms $\|\mathcal{A}_i\|_2$ from the stability characteristics of the linear time-varying system:

$$\dot{x}(t) = A(t)x(t), \quad t \geq 0, \tag{9.7}$$

where the matrix $A(t)$ depends continuously on time.

Recall that the state $x = 0$ is always an equilibrium of the system (9.7). It is the unique equilibrium if $A(t)$ is nonsingular for all $t \geq 0$. The following facts are standard (e.g., [1]).

THEOREM 9.5. *The equilibrium $x = 0$ is stable if and only if the solutions of the linear systems are bounded; that is* $\sup_{t\geq\tau} \|\Phi(t,\tau)\|_2 \stackrel{\text{def}}{=} k(\tau) < \infty \ \forall \tau \geq 0$. *It is uniformly stable if and only if* $\sup_{\tau\geq 0} k(\tau) = \sup_{\tau\geq 0} \sup_{t\geq\tau} \|\Phi(t,\tau)\|_2 \stackrel{\text{def}}{=} \kappa_0 < \infty$. *Finally, it is is exponentially stable if* $\int_0^t \|\Phi(t,\tau)\|_2^2 d\tau \leq C \ \forall t \geq 0$, *where constant C is independent on t.*

Using these stability results, we can obtain some estimates for the growth of $\|\mathcal{A}_i\|_2$.

THEOREM 9.6. *If the equilibrium $x = 0$ is stable and* $k_1 \stackrel{\text{def}}{=} \sup_{t\geq 0} \|B(t)\|_2 < \infty$, *then*

$$\|\mathcal{A}_i\|_2 \quad \leq \quad k_1 \left[\int_0^{t_i} k^2(\tau) d\tau \right]^{1/2}. \tag{9.8}$$

Proof. For all $u \in L^2([0,t_i]R^m)$, we have

$$\begin{aligned}\|\mathcal{A}_i u\| = \| \textstyle\int_0^{t_i} \Phi(t_i,\tau)B(\tau)u(\tau)d\tau\| \ &\leq \ \textstyle\int_0^{t_i} \|\Phi(t_i,\tau)B(\tau)\|_2 \|u(\tau)\|_2 d\tau \\ &\leq \left[\textstyle\int_0^{t_i} \|\Phi(t_i,\tau)\|_2^2 \|B(\tau)\|_2^2 d\tau \cdot \textstyle\int_0^{t_i} \|u(\tau)\|_2^2 d\tau \right]^{1/2} \leq k_1 [\textstyle\int_0^{t_i} k^2(\tau) d\tau]^{1/2} \|u\|_2,\end{aligned}$$

hence inequality (9.8). ∎

This upper bound depends on the growth of integral $\int_0^{t_i} k^2(\tau)d\tau$ with respect to t_i, which can be very fast. Moreover, it can happen that function $k(\cdot)$ does not belong to $L^2([0,t_i])$, in which case bound (9.8) gives no information. However, if we assume uniform stability of the equilibrium $x = 0$, we can get much better bounds.

THEOREM 9.7. *If equilibrium $x = 0$ is uniformly stable and $k_1 < \infty$, then* $\|\mathcal{A}_i\|_2 \leq k_0 k_1 \sqrt{t_i}$.

The proof of this theorem is the same as that of Theorem 9.6. However, we can now ensure a sublinear bound for the growth $\|\mathcal{A}_i\|_2$ with respect to t_i. If we strengthen again the stability assumption, we can obtain an upper bound independent on t_i.

THEOREM 9.8. *Let equilibrium $x = 0$ be exponentially stable and $k_1 < \infty$. Then* $\|\mathcal{A}_i\|_2 \leq k_1 \sqrt{C}$.

Again, this fact can be easily derived from the arguments of the proof of Theorem 9.6. In some cases, we can obtain a computable upper bound for the norm $\|\mathcal{A}_i\|_2$. Recall the following well-known sufficient condition for the global exponential stability.

THEOREM 9.9 (see [1]). *Let the linear system* (9.7) *be time-invariant, and assume there exists a matrix $P = P^T \succ 0$ such that $A^T P + PA \prec 0$. Then equilibrium $x = 0$ is globally exponentially stable.*

Under conditions of this theorem, there exists $\eta_1 > 0$ such that the following LMI

$$A^T P + PA \preceq -\eta_1 P, \quad P = P^T \succ 0,$$

admits a solution. Matrix P and constant η_1 can help us to obtain an explicit upper bound for the norm $\|\mathcal{A}_i\|_2$. Indeed, by definition, $\mathcal{A}_i u$ is the position at time t_i of the point $x(t)$ of the unique trajectory defined by the linear system

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = 0.$$

Therefore, we have $\|x(t_i)\|_2^2 = \langle x(t_i), x(t_i)\rangle \le \frac{\langle Px(t_i), x(t_i)\rangle}{\lambda_{\min}(P)} = \frac{R(t_i)}{\lambda_{\min}(P)}$, where we have introduced function $R(t) \stackrel{\text{def}}{=} \langle Px(t), x(t)\rangle$. Its derivative can be bounded as follows:

$$\begin{aligned}\dot{R}(t) &= \langle P, x(t)\dot{x}(t)^T + \dot{x}(t)x(t)^T\rangle\\ &= \langle P, x(t)(Ax(t)+Bu(t))^T + (Ax(t)+Bu(t))x(t)^T\rangle\\ &= \langle P(Ax(t)+Bu(t)), x(t)\rangle + \langle Px(t), Ax(t)+Bu(t)\rangle\\ &= \langle (PA + A^TP)x(t), x(t)\rangle + 2\langle Px(t), Bu(t)\rangle\\ &\le -\eta_1\langle Px(t), x(t)\rangle + 2\langle Px(t), Bu(t)\rangle \;\le\; \tfrac{1}{\eta_1}\langle PBu(t), Bu(t)\rangle.\end{aligned}$$

Since $x(0) = 0$, we get

$$\begin{aligned}R(t_i) &= \textstyle\int_0^{t_i} \dot{R}(t)dt \;\le\; \frac{1}{\eta_1}\int_0^{t_i}\langle PBu(t), Bu(t)\rangle dt\\ &\le \tfrac{1}{\eta_1}\lambda_{\max}(P)\textstyle\int_0^{t_i}\|Bu(t)\|_2^2\,dt \;\le\; \frac{1}{\eta_1}\lambda_{\max}(P)\,\|B\|_2^2\,\|u\|_2^2.\end{aligned}$$

Hence, $\|\mathcal{A}_i u\|_2^2 \le \frac{\lambda_{\max}(P)}{\eta_1\lambda_{\min}(P)}\|B\|_2^2\|u\|_2^2$, and therefore $\|\mathcal{A}_i\|_2^2 \le \frac{\lambda_{\max}(P)}{\eta_1\lambda_{\min}(P)}\|B\|_2^2$.

If we want to obtain the best upper bound for $\|\mathcal{A}_i\|_2$, we need to solve the following optimization problem in the variables η_1, η_2, η_3, and P:

$$\min\left\{\frac{\eta_3}{\eta_1\eta_2} : A^TP + PA \preceq -\eta_1 P, \quad \eta_2 I \preceq P \preceq \eta_3 I, \quad \eta_1,\eta_2,\eta_3 \ge 0\right\}. \tag{9.9}$$

Although this problem is nonconvex, we can provide an upper bound for its optimal value that is computable with a convex LMI. Indeed, we note that

$$\|\mathcal{A}_i\|_2^2 \le \min\left\{\frac{\eta_3}{\eta_1\eta_2} : A^TP + PA \preceq -\eta_1\eta_3 I, \quad \eta_2 I \preceq P \preceq \eta_3 I, \quad \eta_1,\eta_2,\eta_3 \ge 0\right\}$$

since the feasible set in the right-hand side is smaller than that of (9.9). Furthermore, if we introduce new variables: $\tilde{P} = \frac{P}{\eta_1\eta_3}$, $\tilde{\eta}_2 = \frac{\eta_2}{\eta_1\eta_3}$, and $\tilde{\eta}_3 = \frac{1}{\eta_1}$, we obtain a convex problem that can be solved in polynomial time,

$$\min\left\{\frac{\tilde{\eta}_3^2}{\tilde{\eta}_2} : A^T\tilde{P} + \tilde{P}A \preceq -I, \quad \tilde{\eta}_2 I \le P \le \tilde{\eta}_3 I, \quad \tilde{\eta}_2,\tilde{\eta}_3 \ge 0\right\}.$$

10. Comparison with the literature and conclusion. The subject of this paper can be summarized as the development of an efficient first-order method (obtained using the double smoothing technique) in order to solve partially finite (or finite) convex optimization problems with linear constraints.

Partially finite convex problems have been extensively studied in a theoretical way with duality results, weak constraint qualification [6, 7, 9, 10, 14, 15] and applications, for example, to maximum entropy [5, 8].

On the other hand, it is not the first time that the smoothing technique is used for solving finite-dimensional convex problems with linear constraint by first-order methods. As our approach can also be interesting for solving finite-dimensional problems, we briefly mention these results here, and discuss the differences with our approach.

In [16], the authors consider the case of a conic problem with linear objective function, i.e., $J(u) = \langle c^*, u\rangle$, with $c^* \in U^*$, $T = \{b^*\} \subset V^*$, and $S = \mathcal{L} \subset U$ a closed convex cone. Using the rich duality theory for such kinds of conic problems, they consider a primal-dual approach. The main idea is to reformulate the primal dual optimality conditions $\mathcal{A}^* y + s^* - c^* = 0, \mathcal{A}u - b^* = 0, \langle c^*, u\rangle - \langle b^*, y\rangle = 0, (u, y, s^*) \in \mathcal{L} \times V \times \mathcal{L}^*$ as a nonsmooth convex problem:

$$\overline{f} = \min_{z\in Z}\{f(z) = \|Ez - e\|_*\} = \min_{z\in Z} \max_{\|w\|\leq 1} \langle Ez - e, w\rangle, \tag{10.1}$$

where $\|.\|$ denotes a norm on $U \times V \times R$, $\|\cdot\|_*$ its dual norm, $z = (u, y, s^*)^T$, $e = (c^*, b^*, 0)^T$,

$$E = \begin{pmatrix} 0 & \mathcal{A}^* & I \\ \mathcal{A} & 0 & 0 \\ c^* & -b^* & 0 \end{pmatrix},$$

and $Z = \mathcal{L} \times V \times \mathcal{L}^*$. Applying the smoothing technique to the function f, they are able to find a primal-dual solution z_ϵ such that $\|Ez_\epsilon - e\|_* \leq \epsilon$ in $O\left(\frac{1}{\epsilon}\right)$. This primal-dual approach does not work in our case for two reasons. First, primal-dual optimality conditions cannot be expressed as a linear system $Ez = e$ (subject to a conic constraint). Furthermore, we do not want to work in the primal-dual space but preferably in the dual one due to our asymmetry assumption (the problem (10.1) can be infinite-dimensional in our framework).

The approach considered in [2] is more comparable to what we are doing. Their problem class is composed of problems of the form (3.1) with J a (not necessarily smooth) convex function, $S = U$ and $T = \mathcal{K} - b$, where $\mathcal{K}$ is a closed convex cone. Dualizing the constraint $\mathcal{A}u+b \in \mathcal{K}$, they obtain a dual problem with conic constraint: $\max_{z\in\mathcal{K}^*} g(z)$, where $g(z) = \inf_u J(u) - \langle \mathcal{A}u + b, z\rangle$. They apply the smoothing technique to the dual objective function and compare different optimal first-order methods of smooth convex optimization for solving the smoothed dual problem. They are able to solve the dual problem with accuracy ϵ in $O\left(\frac{1}{\epsilon}\right)$ iteration. To reconstruct a nearly optimal primal solution u_ϵ from a nearly optimal dual solution z_ϵ, they suggest choosing u_ϵ as the minimizer of the optimization subproblem defining the smoothed dual objective function at the point z_ϵ. However, this suggestion is not supported by the analysis of the convergence rates (see our discussion at the end of section 5.1).

In the framework of separable convex problems, the smoothing technique has also been applied in [17] to convex problems with linear coupling constraint. Dualizing the coupling constraint, the authors obtain a dual objective function that can be computed in a separable way. Applying a simple smoothing to this dual objective function, they obtain a smooth dual objective function keeping the separability structure. Here also, this approach allows them to solve the dual problem with accuracy ϵ in $O\left(\frac{1}{\epsilon}\right)$ iterations. Averaging of the minimizers of the subproblems defining the smoothed dual objective function at the different dual iterates is proposed to reconstruct a primal solution. It is proved that the quality of this primal solution is also of order ϵ. It also depends on the norm of the dual optimal solution (which is typically unknown).

The approach considered in this paper also allows us to exploit the separability structure of decomposable problems with linear coupling constraints (see the two examples given in section 4). In our work, we apply a double smoothing that gives us a possibility to reconstruct more easily a nearly optimal primal solution from a nearly optimal dual solution, without averaging. The price that we pay for this simplicity is a logarithmic term $\log\left(\frac{1}{\epsilon}\right)$ in the complexity. For the level of accuracy we are interested in, the logarithmic factor is not distinguishable from an absolute constant. Furthermore, whereas we also use in our analysis the norm of the dual optimal solution, we are able to provide an explicit upper bound for this quantity.

More generally, to the best of our knowledge, this work is the first one where the smoothing technique is applied for solving infinite-dimensional problems. In particular, we considered in section 9 optimal control problems governed by a system of linear differential equations with the constraint that the trajectory crosses some convex set at certain moments in time.

Acknowledgments. The authors would like to thank the anonymous referee for very useful comments that have improved the quality of this paper and the practical applicability of the method.

REFERENCES

[1] P. J. Antsaklis and A. N. Michel, *Linear Systems*, Birkhäuser Boston, Boston, 2006.
[2] S. Becker, E. Candès and M. Grant, *Templates for convex cone problems with applications to sparse signal recovery*, Math. Program. Comput., 3 (2011), pp. 165–218.
[3] P. Bernhard and A. Rapaport, *On a theorem of Danskin with an application to a theorem of Von Neumann-Sion*, Nonlinear Anal., 24 (1995), pp. 1163–1181.
[4] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
[5] J. M. Borwein and A. S. Lewis, *Duality relationship for entropy-like minimization problems*, SIAM. J. Control Optim., 29 (1991), pp. 325–338.
[6] J. M. Borwein and A. S. Lewis, *Partially finite convex programming,* I. *Quasi relative interiors and duality theory*, Math. Programming, 57 (1992), pp. 15–48.
[7] J. M. Borwein and A. S. Lewis, *Partially finite convex programming,* II. *Explicit lattice models,* Math. Programming, 57 (1992), pp. 49–83.
[8] J. M. Borwein and A. S. Lewis, *Partially finite programming in L_1 and the existence of maximum entopy estimates*, SIAM. J. Optim., 3 (1993), pp. 248–267.
[9] J. M. Borwein and H. Wolkowicz, *A simple constraint qualification in infinite-dimensional programming,* Math. Programming, 35 (1986), pp. 83–96.
[10] R. I. Bot, S-M. Grad, and G. Warka, *New regularity conditions for Lagrange and Fenchel-Lagrange duality in infinite dimensional spaces*, Math. Inequal. Appl., 12 (2009), pp. 171–189.
[11] J. M. Danskin, *The theory of max-min and its application to weapons allocation problems*, Springer-Verlag, New York, 1967.
[12] D. E. Kirk, *Optimal Control: An Introduction*, Dover, New York, 2004.
[13] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*. Wiley Interscience, New York, 1972.
[14] V. Jeyakumar, N. Dinh and G.M. Lee, *A New closed cone constraint qualification for convex optimization*, Applied Mathematics Report AMR 04/08, University of New South Wales, Kensington, Australia, 2004.
[15] V. Jeyakumar and H. Wolkowicz, *Generalizations of Slater's constraint qualification for infinite convex programs,* Math. Programming, 57 (1992), pp. 85–101.
[16] G. Lan, Z. Lu and R. D. C Monteiro, *Primal-dual first-order methods with $O(\frac{1}{\epsilon})$ iteration-complexity for cone programming*, Math. Program., 126 (2011), pp. 1–29.
[17] I. Necoara and J. A. K. Suykens, *Application of a smoothing technique to decomposition in convex optimization*, IEEE Trans. Automat. Control, 53 (2008), pp. 2674–2679.
[18] A. Nemirovsky and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley & Sons, New York, 1983.
[19] Yu. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Boston, 2004.
[20] Yu. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., Ser. A, 103 (2005), pp. 127–152.
[21] Yu. Nesterov, *Excessive gap technique in nonsmooth convex minimization*, SIAM J. Optim., 16 (2005), pp. 235–249.
[22] Yu. Nesterov, *Smoothing technique and its applications in semidefinite optimization,* Math. Program. Ser. A, 110 (2007), pp. 245–259.
[23] Yu. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program., Ser. B, 120 (2009), pp. 221–259.
[24] Yu. Nesterov, *A method for solving the convex programming problem with convergence rate $O(\frac{1}{k^2})$*, Dokl. Akad SSSR, 269 (1983), pp. 543–547.

[25] Yu. Nesterov, *On an approach to the construction of optimal methods of minimization of smooth convex function*, Èkonom. i. Mat. Metody (in Russian), 24 (1988), pp. 509–517.
[26] E. Zeidler, *Nonlinear Functional Analysis and its Application,* III*: Variational Methods and Optimization*, Springer-Verlag, New York, 1985.