Computational Optimization and Applications (2019) 74:821–850
https://doi.org/10.1007/s10589-019-00124-7

Smoothing algorithms for computing the projection onto a Minkowski sum of convex sets

Xiaolong Qin[1,2] · **Nguyen Thai An**[2,3]

Received: 13 March 2018 / Published online: 22 August 2019

Abstract
In this paper, the problem of computing the projection, and therefore the minimum distance, from a point onto a Minkowski sum of general convex sets is studied. Our approach is based on Nirenberg's minimum norm duality theorem and Nesterov's smoothing techniques. It is shown that the projection onto a Minkowski sum of sets can be represented as the sum of points on constituent sets so that, at these points, all of the sets share the same normal vector which is the negative of the dual solution. For numerically solving the problem, the most suitable algorithm is the one suggested by Gilbert (SIAM J Control 4:61–80, 1966). This algorithm has been widely used in collision detection and path planning in robotics. However, a main drawback of this method is that in some cases, it turns to be very slow as it approaches the solution. In this paper we proposed NESMINO whose $O\left(\frac{1}{\sqrt{\epsilon}}\ln(\frac{1}{\epsilon})\right)$ complexity bound is better than the worst-case complexity bound of $O(\frac{1}{\epsilon})$ of Gilbert's algorithm.

Keywords Minimum norm problem · Minkowski sum of sets · Gilbert's algorithm · Nesterov's smoothing technique · Fast gradient method · SAGA

Mathematics Subject Classification Primary 49J52 · 49M29 · Secondary 90C30

✉ Xiaolong Qin
qxlxajh@163.com

Nguyen Thai An
thaian2784@gmail.com

1 Department of Mathematics, Hangzhou Normal University, Hangzhou, China

2 Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China

3 Institute of Research and Development, Duy Tan University, Danang, Vietnam

1 Introduction

Let A and B be two subsets in $\mathbb{R}^n$. Recall that the *Minkowski sum* of these sets is defined by

$$A + B := \{a + b : \ a \in A, b \in B\}.$$

The case of more than two sets is defined in the same way by induction. Note that if all the sets A_i for $i = 1, \ldots, m$ are convex, then every linear combination of these sets, $\sum_{i=1}^{m} \lambda_i A_i$ with $\lambda_i \in \mathbb{R}$ for $i = 1, \ldots, m$, is also convex. The Euclidean distance function associated with a subset Q is defined by

$$d(q; Q) := \inf\{\|x - q\| : \ x \in Q\},$$

where $\|\cdot\|$ is the Euclidean norm. The optimization problem we are concerned with in this paper is the following *minimum norm problem*

$$d\left(0; \sum_{i=1}^{p} T_i(\Omega_i)\right) := \min\left\{\|x\| : \ x \in \sum_{i=1}^{p} T_i(\Omega_i)\right\}, \tag{1.1}$$

where Ω_i, for $i = 1, \ldots, p$, are nonempty convex compact sets in $\mathbb{R}^m$ and $T_i : \mathbb{R}^m \to \mathbb{R}^n$ are affine mappings satisfying

$$T_i(y) = A_i y + a_i, \quad y \in \mathbb{R}^m,$$

where A_i, for $i = 1, \ldots, p$, are $n \times m$ matrices and a_i, for $i = 1, \ldots, p$, are given points in $\mathbb{R}^n$. Since $\sum_{i=1}^{p} T_i(\Omega_i)$ is closed and convex, and the norm under consideration is Euclidean, (1.1) has a unique solution, which is the projection from the origin onto $\sum_{i=1}^{p} T_i(\Omega_i)$. We denote this solution by x^* throughout this paper.

We assume in problem (1.1) that each set Ω_i is simple enough so that the corresponding projection operator P_{Ω_i} is easy to compute. It is worth noting that there hasn't been an algorithm for finding the Minkowski sum of general convex sets except for the cases of balls and polytopes. Moreover, the projection onto the Minkowski sum of closed convex sets is generally not equal to the sum of the projections onto constituent sets. A complete answer to the question "When is the sum of projectors also a projector?" has been provided very recently by Bauschke et al. [1].

Minimum norm problems for the case of polytopes have been well studied in the literature from both theoretical and numerical point of view; see e.g., [14,27,39] and the references therein. The most suitable algorithm for solving (1.1) is perhaps the one suggested by Gilbert [17]. The original Gilbert's algorithm was devised for solving the minimum norm problem associated with just one convex compact set. The algorithm does not require the explicit projection operator of the given set. Instead, it requires in each step the computation of the support point of the set along with a certain direction. By observation that for a given direction, support point of a Minkowski sum of sets can be represented in terms of support points of constituent sets, Gilbert's algorithm

thus can be applied for general case of (1.1). Gilbert's algorithm can be analyzed from geometrical point of views and is easy to implement. However, a serious problem in the algorithm is that, in some cases, it loops infinitely and is very slow as it approaches the solution; see [6,23,26]. Following [17], Gilbert's algorithm is a descent method that generates a sequence $\{z_k\}$ satisfying $\|z_k\|$ converges downward to $\|x^*\|$ within $O(\frac{1}{k})$ iterations.

Another effective algorithm for distance computation between two convex objects is the Gilbert–Johnson–Keerthi (GJK) algorithm proposed in [18] and its enhancing versions [3,5,19]. The original GJK algorithm was just restricted to compute the distance between objects which can be approximately represented as convex polytopes. In order to reduce the error of the polytope approximations in finding the minimum distance, Gilbert and Fo [19] modified the original GJK to handle general convex objects. The new modified algorithm is based on Gilbert's algorithm and has the same bound on number of iterations.

To illustrate another utility of Minkowski projection, we consider the problem of minimizing a function f on an intersection $\bigcap_{i=1}^{k} C_i$ of conic constraints. It is known that the indicator function of a closed convex cone C coincides with the support function of its polar; i.e., $\delta_C = \sigma_{C^\circ}$. The problem can be reformulated as

$$\min_{x\in\mathbb{R}^n} f(x) + \sigma_{C_1^\circ}(x) + \cdots + \sigma_{C_k^\circ}(x) = \min_{x\in\mathbb{R}^n} f(x) + \sigma_{C_1^\circ+\cdots+C_k^\circ}(x). \tag{1.2}$$

The equality above holds since support functions are additive over Minkowski sums; see [20]. This model includes the classical lasso [36] and its generalizations such as group lasso [41,42], constrained lasso [15,37]. Standard tools for solving this problems are proximal-type algorithms that require computing the proximal operator of the function σ_Q, where $Q = C_1^\circ + \cdots + C_k^\circ$ is the Minkowski sum of the polar sets. As a consequence of the Moreau decomposition (see [2, Theorem 6.46]), the following relationship allows us to compute this operator in terms of the Euclidean projection onto the Minkowski sum:

$$\mathrm{prox}_{\lambda\sigma_Q}(x) = x - \lambda P_Q(\lambda^{-1}x).$$

Contributions The contributions of this paper are two-fold: On the theoretical side we show that each projection onto a Minkowski sum of convex sets can be represented as the sum of points on constituent sets so that at these points, all the sets share the same normal vector. This result is obtained by investigating the relationship between solutions of (1.1) and its Fenchel duality. We also give conditions for the uniqueness of solution of primal and dual problem. On the numerical side, we proposed NESMINO that is based on the smoothing technique developed by Nesterov [30,31]. To this end, we first approximate the dual objective function by a smooth and strongly convex function and solve this smooth problem via a fast gradient scheme. After that we show how an approximate solution for the primal problem, i.e., an approximation for the projection of the origin, can be reconstructed from the dual iterative sequence. Our algorithm has the complexity bound of $O\left(\frac{1}{\sqrt{\epsilon}}\ln(\frac{1}{\epsilon})\right)$ that is better than the worst-case complexity $O(\frac{1}{\epsilon})$ of Gilbert's algorithm. Moreover, the algorithm also provides

elements on each constituent sets such that their sum is equal to the projection x^*. When the number of sets in the Minkowski sum is large, we introduce SAGA-NESMINO which is comparable to Gilbert's algorithm from both running time and accuracy.

The rest of the paper is organized as follows. In Sect. 2, we provide tools of convex analysis that are widely used in the sequel. The Nesterov's smoothing technique and fast gradient method are recalled in Sect. 3. In Sect. 4, we state some duality results concerning the minimum norm problems. Section 5 is devoted to an overview of Gilbert's algorithm. Section 6 is the main part of the paper devoted to develop a smoothing algorithms for solving (1.1). Some numerical experiments are also provided in this section.

2 Tools of convex analysis

Let $\langle\cdot,\cdot\rangle$ be the inner product associated with Euclidean norm $\|\cdot\|$ in $\mathbb{R}^n$. An extended real-valued function $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ is said to be convex if

$$f((1-\lambda)x+\lambda y)\le(1-\lambda)f(x)+\lambda f(y)$$

for all $x, y\in\mathbb{R}^n$ and $\lambda\in(0,1)$. We say that f is *strongly convex* with modulus γ if $f-\frac{\gamma}{2}\|\cdot\|^2$ is a convex function. Let Q be a subset of $\mathbb{R}^n$, the support function of Q is defined by

$$\sigma_Q(u):=\sup\{\langle u,x\rangle:\ x\in Q\},\ u\in\mathbb{R}^n. \tag{2.1}$$

It follows directly from the definition that σ_Q is positive homogeneous and subadditive. The set-valued mapping $S_Q:\mathbb{R}^n\rightrightarrows\mathbb{R}^n$ defined by

$$S_Q(u)=\{x\in Q:\ \langle u,x\rangle=\sigma_Q(u)\},\ u\in\mathbb{R}^n$$

is called the *support point mapping* of Q. If Q is compact, then $\sigma_Q(u)$ is finite and $S_Q(u)\neq\emptyset$ for all $u\in\mathbb{R}^n$.

In order to study minimum norm problem in which the Euclidean distance is replaced by distances generated by different norms, we consider a more general setting. Let F be a closed, bounded and convex set of $\mathbb{R}^n$ that contains the origin as an interior point. The *minimal time function* associated with the dynamic set F and the target set Q is defined by

$$T_F(x;Q):=\inf\{t\ge0:\ (x+tF)\cap Q\neq\emptyset\}. \tag{2.2}$$

The minimal time function (2.2) can be expressed as

$$T_F(x;Q)=\inf\{\rho_F(\omega-x):\ \omega\in Q\}, \tag{2.3}$$

where $\rho_F(x):=\inf\{t\ge0:\ x\in tF\}$ is the Minkowski function associated with F. Moreover, $T_F(\cdot,Q)$ is convex if and only if Q is convex; see [28]. We denote the set of *generalized projection* from x to Q by

$$\Pi_F(x;\, Q) := \{q \in Q : \rho_F(q - x) = T_F(x;\, Q)\}.$$

Note that, if F is the closed unit ball generated by some norm $|||\cdot|||$ on $\mathbb{R}^n$, then we have $\rho_F = |||\cdot|||$, $\sigma_F = |||\cdot|||_*$ and $T_F(\cdot,\, Q)$ reduces to the ordinary distance function

$$d(x;\, Q) = \inf\{|||\omega - x||| :\ \omega \in Q\}, \quad x \in \mathbb{R}^n.$$

The set $\Pi_F(x;\, Q)$ in this case is denoted by $\Pi(x;\, Q) := \{q \in Q :\ d(x;\, Q) = |||q - x|||\}$. When $|||\cdot|||$ is Euclidean norm, we simply use the notation $P_Q(x)$ instead. If Q is a nonempty closed convex set, then the Euclidean projection $P_Q(x)$ is a singleton for every $x \in \mathbb{R}^n$.

The following results whose proof can be found in [20] allow us to represent support functions of general sets in term of the support functions of one or more simpler sets.

Lemma 2.1 *Consider the support function* (2.1). *Let Ω, Ω_1, Ω_2 be subsets of $\mathbb{R}^m$ and $T : \mathbb{R}^m \to \mathbb{R}^n$ satisfying $T(x) = Ax + a$ be an affine transformation, where A is an $n \times m$ matrix and $a \in \mathbb{R}^n$. The following assertions hold:*

(i) $\sigma_\Omega = \sigma_{\mathrm{cl}\,\Omega} = \sigma_{\mathrm{co}\,\Omega} = \sigma_{\overline{\mathrm{co}}\,\Omega}$.
(ii) $\sigma_{\Omega_1+\Omega_2}(u) = \sigma_{\Omega_1}(u) + \sigma_{\Omega_2}(u)$ *and* $\sigma_{\Omega_1-\Omega_2}(u) = \sigma_{\Omega_1}(u) + \sigma_{\Omega_2}(-u)$, *for all* $u \in \mathbb{R}^m$.
(iii) $\sigma_{T(\Omega)}(v) = \sigma_\Omega(A^\top v) + \langle v, a\rangle$, *for all* $v \in \mathbb{R}^n$.

A subset $\Omega \in \mathbb{R}^m$ is said to be *strictly convex* if $tu + (1-t)v \in \mathrm{int}(\Omega)$ whenever $u, v \in \Omega$, $u \neq v$ and $t \in (0, 1)$. The proof of the following result is straightforward.

Lemma 2.2 *Let Ω, Ω_1, Ω_2 be convex compact subsets of $\mathbb{R}^m$ and $T : \mathbb{R}^m \to \mathbb{R}^n$ be an affine transformation satisfying $T(x) = Ax + a$, where A is an $n \times m$ matrix and $a \in \mathbb{R}^n$. The following assertions hold:*

(i) $S_{\Omega_1+\Omega_2}(u) = S_{\Omega_1}(u) + S_{\Omega_2}(u)$, *for all* $u \in \mathbb{R}^m$.
(ii) $S_{T(\Omega)}(v) = T\left(S_\Omega(A^\top v)\right) = A\left(S_\Omega(A^\top v)\right) + a$, *for all* $v \in \mathbb{R}^n$.
(iii) *If suppose further that Ω is a strictly convex set, then $S_\Omega(u)$ is a singleton for any* $u \in \mathbb{R}^m \setminus \{0\}$.

The *Fenchel conjugate* of a convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is defined by

$$f^*(v) := \sup\{\langle v, x\rangle - f(x) :\ x \in \mathbb{R}^n\},\ \ v \in \mathbb{R}^n.$$

If f is proper and lower semicontinuous, then $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is also a proper, lower semicontinuous convex function. From the definition, support function σ_Q is the Fenchel conjugate of the *indicator function* δ_Q of Q which is defined by $\delta_Q(x) = 0$ if $x \in Q$ and $\delta_Q(x) = +\infty$ otherwise.

The polar of a subset $E \subset \mathbb{R}^n$ is the set $E^\circ = \{u \in \mathbb{R}^n :\ \sigma_E(u) \leq 1\}$. When E is the closed unit ball of a norm $|||\cdot|||$, then E° is the closed unit ball of the corresponding dual norm $|||\cdot|||_*$. Some basis properties of the polar set can be found in [38, Proposition 1.23]. If F is a closed convex and bounded set that contains the origin in its interior, then F° is also a closed convex and bounded set with $0 \in \mathrm{int}(F^\circ)$. Moreover, $\rho_F = (\delta_{F^\circ})^* = \sigma_{F^\circ}$ is a Lipschitz function with modulus $\|F^\circ\| := \sup\{\|x\| :\ x \in F^\circ\}$.

Let us recall below the Fenchel duality theorem which plays an important role in the sequel. We denote the set of points where a function $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is finite and continuous by $\text{cont}(g)$.

Theorem 2.1 (See [4, Theorem 3.3.5]) *Given functions $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, and a linear mapping $A : \mathbb{R}^m \to \mathbb{R}^n$, the weak duality inequality*

$$\inf_{x \in \mathbb{R}^m} \{f(x) + g(Ax)\} \geq \sup_{u \in \mathbb{R}^n} \{-f^*(A^*u) - g^*(-u)\}$$

holds. If furthermore f and g are convex and satisfy the following condition $A(\text{dom } f) \cap \text{cont}(g) \neq \emptyset$, then the equality holds and the supremum is attained if it is finite.

3 Nesterov's smoothing technique and fast gradient method

In a celebrated work [32], Nesterov introduced a fast first-order method for solving convex smooth problems in which the objective functions have Lipschitz continuous gradients. In contrast to the complexity bound of $O(\frac{1}{\epsilon})$ possessed by the classical gradient descent method, Nesterov's method gives a complexity bound of $O(\frac{1}{\sqrt{\epsilon}})$, where ϵ is the desired accuracy for the objective function.

When the problem under consideration is nonsmooth in which the objective function has an explicit max-structure as follows

$$f(u) := \max\{\langle Au, x\rangle - \phi(x) : \ x \in Q\}, \ u \in \mathbb{R}^n,$$

where A is an $m \times n$ matrix and ϕ is a continuous convex function on a compact set Q of $\mathbb{R}^m$, in order to overcome the complexity bound $O(\frac{1}{\epsilon^2})$ of the subgradient method, Nesterov [30] made use of the special structure of f to approximate it by a function with Lipschitz continuous gradient and then applied a fast gradient method to minimize the smooth approximation. With this combination, we can solve the original non-smooth problem up to accuracy ϵ in $O(\frac{1}{\epsilon})$ iterations. To this end, let d be a continuous strongly convex function on Q. Let μ be a positive number called a *smooth parameter*. Define

$$f_\mu(u) := \max\{\langle Au, x\rangle - \phi(x) - \mu d(x) : \ x \in Q\}. \tag{3.1}$$

Since $d(x)$ is strongly convex, problem (3.1) has a unique solution. The following statement is a simplified version of [30, Theorem 1].

Theorem 3.1 *The function f_μ in (3.1) is well defined and continuously differentiable on $\mathbb{R}^n$. The gradient of the function is*

$$\nabla f_\mu(u) = A^\top x_\mu(u),$$

where $x_\mu(u)$ is the unique element of Q such that the maximum in (3.1) is attained. Moreover, ∇f_μ is a Lipschitz function with the Lipschitz constant $\ell_\mu = \dfrac{1}{\mu\sigma_1}\|A\|^2$, and

$$f_\mu(u) \le f(u) \le f_\mu(u) + \mu D \quad \forall u \in \mathbb{R}^n,$$

where $D := \max\{d(x) : \ x \in Q\}$.

For the reader's convenience, we conclude this section with a presentation of the simplest optimal method for minimizing smooth and strongly convex functions; see [31] and the references therein. Let $g : \mathbb{R}^n \to \mathbb{R}$ be strongly convex with parameter $\gamma > 0$ and its gradient be Lipschitz continuous with constant $L \ge \gamma$. The fast gradient method for solving problem $g^* = \inf\{g(u) : u \in \mathbb{R}^n\}$ is outlined as follows:

Fast Gradient Method
INITIALIZE: $u_0 = v_0 \in \mathbb{R}^n$.
Set $k = 0$.
Repeat the following
Set $u_{k+1} := v_k - \frac{1}{L}\nabla g(v_k)$
Set $v_{k+1} := u_{k+1} + \frac{\sqrt{L}-\sqrt{\gamma}}{\sqrt{L}+\sqrt{\gamma}}(u_{k+1} - u_k)$
Set $k := k + 1$
Until a stopping criterion is satisfied.

4 Duality for minimum norm problems

In this section, we are in a position to give some duality results concerning minimum norm problem (1.1). Let us first recall the duality theorem originally stated by Nirenberg [33].

Theorem 4.1 *(Minimum norm duality theorem) Given $\bar{x} \in \mathbb{R}^n$ and let $d(\cdot; \Omega)$ be the distance function to a nonempty closed convex set Ω associated with some norm $|||\cdot|||$ on $\mathbb{R}^n$. Then*

$$d(\bar{x}; \Omega) = \max\{\langle u, \bar{x}\rangle - \sigma_\Omega(u) : \ |||u|||_* \le 1\},$$

where the maximum on the right is achieved at some $\bar{u}$. Moreover, if $\bar{w} \in \Pi(\bar{x}; \Omega)$, then $-\bar{u}$ is aligned with $\bar{w} - \bar{x}$, i.e., $\langle -\bar{u}, \bar{w} - \bar{x}\rangle = |||\bar{u}|||_* . |||\bar{w} - \bar{x}|||$.

According to this theorem, the minimum distance from a point to a convex set is equal to the maximum of the distance from the point to hyperplanes separating the point and the set; see Fig. 1. A standard proof of this theorem can be found in [25, p. 136]. We also refer the readers to the recent paper [10] for more types of minimum norm duality theorems concerning the width and the length of symmetrical convex bodies.

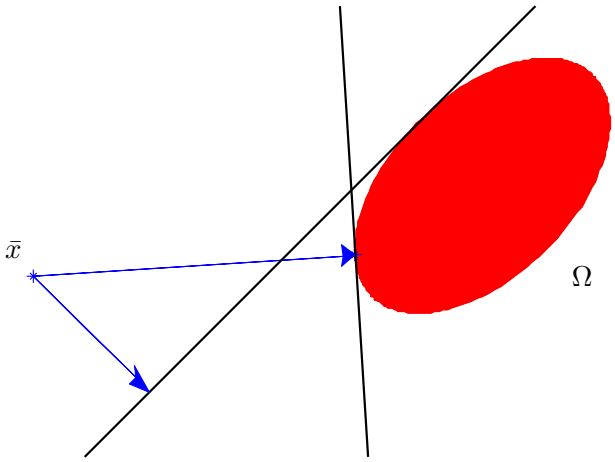


Fig. 1 An illustration of the minimum norm duality theorem

Lemma 4.1 *Let Q be a nonempty closed subset of $\mathbb{R}^n$. Then the generalized projection $\Pi_F(x; Q)$ is nonempty for any $x \in \mathbb{R}^n$.*

Proof Since F is a closed bounded and convex set that contains the origin as an interior point, $0 \leq T_F(x; Q) < +\infty$ for all $x \in \mathbb{R}^n$ and the following number exists

$$R = \sup\{r : \mathbb{B}(0; r) \subset F^\circ\} < +\infty.$$

Then we have $\rho_F(x) = \sigma_{F^\circ}(x) \geq R\|x\|$ for all $x \in \mathbb{R}^n$. Fix $x \in \mathbb{R}^n$. For each $n \in \mathbb{N}$, from (2.3) there exists $w_n \in Q$, such that

$$T_F(x; Q) \leq \rho_F(w_n - x) < T_F(x; Q) + \frac{1}{n}. \tag{4.1}$$

It follows from (4.1) and triangle inequality that

$$R\|w_n\| \leq R\left(\|w_n - x\| + \|x\|\right) \leq \rho_F(w_n - x) + R\|x\| \leq T_F(x; Q) + 1 + R\|x\|$$

for all n. Thus the sequence $\{w_n\}$ is bounded. We can take a subsequence $\{w_{k_n}\}$ that converges to a point $\bar{w} \in Q$ due to the closedness of Q. By taking the limit both sides of (4.1) and using the continuity of the Minkowski function, we can conclude that $\bar{w} \in \Pi_F(x; Q)$. □

Theorem 4.1 is in fact a consequence of the Fenchel duality theorem which is used to prove the following extension for minimal time functions.

Theorem 4.2 *The generalized distance $T_F(0; A(\Omega))$ from the origin $0_{\mathbb{R}^n}$ to the image $A(\Omega)$ of a nonempty closed convex set $\Omega \subset \mathbb{R}^m$ under a linear mapping $A : \mathbb{R}^m \to \mathbb{R}^n$ can be computed by*

$$T_F(0; A(\Omega)) := \inf\{\rho_F(Aw) : w \in \Omega\} = \max\left\{-\sigma_\Omega(-A^\top u) : u \in F^\circ\right\},$$

where the maximum on the right is achieved at some $\bar{u} \in F^\circ$. *If* $A\bar{w} \in \Pi_F(0; A(\Omega))$ *is a projection from the origin onto* $A(\Omega)$, *then*

$$\langle A\bar{w}, \bar{u}\rangle = \sigma_{F^\circ}(A\bar{w}) = -\sigma_\Omega(-A^\top \bar{u}).$$

Proof Applying Theorem 2.1 for $g = \rho_F$ and $f = \delta_\Omega$, the following qualification condition holds

$$A\mathrm{dom}\, f \cap \mathrm{cont}(g) = A(\Omega) \cap \mathbb{R}^n = A(\Omega) \neq \emptyset,$$

where $\mathrm{cont}(g) = \mathbb{R}^n$ is due to the fact that ρ_F is a continuous function on $\mathbb{R}^n$. It follows that

$$\begin{aligned} T_F(0; A(\Omega)) &= \inf\{\delta_\Omega(x) + \rho_F(Ax) : \ x \in \mathbb{R}^n\} \\ &= \sup\{-(\delta_\Omega)^*\left(A^\top u\right) - (\rho_F)^*(-u) : \ u \in \mathbb{R}^n\} \\ &= \sup\{-\sigma_\Omega\left(A^\top u\right) - \delta_{F^\circ}(-u) : \ u \in \mathbb{R}^n\} \\ &= \sup\{-\sigma_\Omega\left(-A^\top u\right) - \delta_{F^\circ}(u) : \ u \in \mathbb{R}^n\} \\ &= \sup\{-\sigma_\Omega\left(-A^\top u\right) : \ u \in F^\circ\}, \end{aligned}$$

and the supremum is attained because $T_F(0; A(\Omega))$ is finite. If the supremum on the right is achieved at some $\bar{u} \in F^\circ$ and the infimum on the left is achieved at some $\bar{w} \in \Omega$, then

$$\begin{aligned} \sigma_{F^\circ}(A\bar{w}) = \rho_F(A\bar{w}) = T_F(0, A(\Omega)) &= \max\{-\sigma_\Omega\left(-A^\top u\right) : \ u \in F^\circ\} \\ &= -\sigma_\Omega(-A^\top \bar{u}). \end{aligned}$$

Since $\bar{w} \in \Omega$, we also have

$$\langle -A^\top \bar{u}, \bar{w}\rangle \leq \sigma_\Omega\left(-A^\top \bar{u}\right) = -\sigma_{F^\circ}(A\bar{w}).$$

This implies that $\langle A^\top \bar{u}, \bar{w}\rangle \geq \sigma_{F^\circ}(A\bar{w})$. On the other hand, $\sigma_{F^\circ}(A\bar{w}) \geq \langle A\bar{w}, \bar{u}\rangle = \langle A^\top \bar{u}, \bar{w}\rangle$, because $\bar{u} \in F^\circ$. Thus, $\langle A\bar{w}, \bar{u}\rangle = \sigma_{F^\circ}(A\bar{w})$. This completes the proof. □

Note that, given a closed set Ω, the set $A(\Omega)$ need not to be closed and therefore, we cannot use the min to replace the inf in the primal problem in Theorem 4.2.

Proposition 4.1 *Let* Q *be a nonempty, closed convex subset of* $\mathbb{R}^n$. *The following holds*

$$T_F(0; Q) := \min\{\rho_F(q) : \ q \in Q\} = \max\{-\sigma_Q(-u) : \ u \in F^\circ\}. \tag{4.2}$$

If the maximum on the right is achieved at $\bar{u} \in F^\circ$ and the infimum on the left is attained at $\bar{q} \in Q$, then

$$\langle \bar{q}, \bar{u} \rangle = \sigma_{F^\circ}(\bar{q}) = -\sigma_Q(-\bar{u}). \tag{4.3}$$

If $F = \mathbb{B}$ is the Euclidean closed unit ball, then the projection $\bar{q}$ exists uniquely and

$$d(0; Q) := \min\{\|q\| : \ q \in Q\} = \max\{-\sigma_Q(-u) : \ u \in \mathbb{B}\}.$$

If suppose further that $0 \notin Q$, then $\frac{\bar{q}}{\|\bar{q}\|}$ is the unique solution of the dual problem.

Proof The first assertion is a direct consequence of Theorem 4.2 with $\Omega = Q$ and A is the identity mapping of $\mathbb{R}^n$. Note that, by Lemma 4.1, the infimum is attained here. When F is the Euclidean ball, the minimal time function reduces to the Euclidean distance function and therefore the projection $\bar{q} = P_Q(0)$ exists uniquely. If $0 \notin Q$, then $\bar{q} \neq 0$. Moreover, we have $\langle -\bar{q}, x - \bar{q} \rangle \leq 0$ for all $x \in Q$. This implies,

$$\left\langle -\frac{\bar{q}}{\|\bar{q}\|}, x \right\rangle \leq -\|\bar{q}\|, \text{ for all } x \in Q.$$

Hence $\sigma_Q\left(-\frac{\bar{q}}{\|\bar{q}\|}\right) \leq -\|\bar{q}\| = -d(0; Q)$. This means that $\frac{\bar{q}}{\|\bar{q}\|}$ is a solution of the following dual problem

$$d(0; Q) = \max\{-\sigma_Q(-u) : \ u \in \mathbb{B}\}.$$

Using (4.3), any dual solution $\bar{u}$ must satisfy $\bar{u} \in S_{F^\circ}(\bar{q})$. Since $F = \mathbb{B}$, we have $F^\circ = \mathbb{B}$ is a strictly convex set. Thus, by Lemma 2.2(iii), $\bar{u} = \frac{\bar{q}}{\|\bar{q}\|}$ is the unique solution of dual problem. The proof is now complete. □

From (4.3), for any primal-dual pair $(\bar{q}, \bar{u})$, we have the following relationship

$$\bar{u} \in S_{F^\circ}(\bar{q}) \quad \text{and} \quad \bar{q} \in S_Q(-\bar{u}). \tag{4.4}$$

This observation seems to be useful from numerical point of view in the sense that if a dual solution $\bar{u}$ is found exactly, then a primal solution $\bar{q}$ can be obtained by taking a support point in $S_Q(-\bar{u})$. However, for a general convex set Q, the set $S_Q(-\bar{u})$ might contain more than one point and there might be some points in this set which is not a desired primal solution; see Fig. 2. Thus, the above task is possible when $S_Q(-\bar{u})$ is a singleton.

When the distance function under consideration is non-Euclidean, the primal problem may have infinitely many solutions and we may not recover a dual solution from a primal one $\bar{q}$ by setting $\frac{\bar{q}}{\|\bar{q}\|}$ as in the Euclidean case.

Example 4.1 In $\mathbb{R}^2$, consider the problem of finding the projection onto the set $Q = \{x \in \mathbb{R}^2 : \ 2 \leq x_1 \leq 5 \text{ and } 1 \leq x_2 \leq 4\}$ in which the distance function generated by the ℓ_∞-norm. In this case, $F = \{x \in \mathbb{R}^2 : \ \max\{|x_1|, |x_2|\} \leq 1\}$ and $F^\circ = \{x \in$

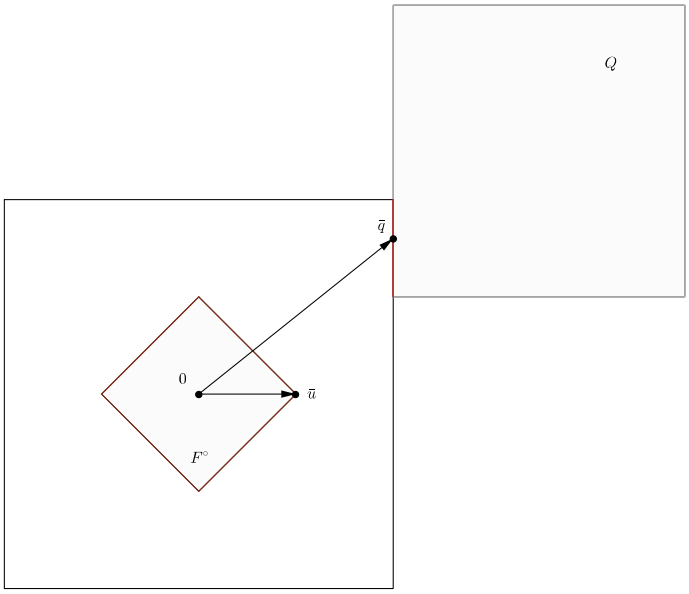


Fig. 2 A minimum norm problem with non-Euclidean distance

$\mathbb{R}^2 : \ |x_1| + |x_2| \leq 1\}$ and we have $T_F(0; Q) = 2$. The primal problem in (4.2) has the solution set $\Pi_F(0; Q) = \{x \in \mathbb{R}^2 : \ x_1 = 2 \text{ and } 1 \leq x_2 \leq 2\}$ and the corresponding dual problem has a unique solution $\bar{u} = (1, 0)$. We can see that, for any primal solution $\bar{q}$, the element $\frac{\bar{q}}{\|\bar{q}\|} \neq \bar{u}$. Thus, $\frac{\bar{q}}{\|\bar{q}\|}$ is not a dual solution; see Fig. 2.

We now give a sufficient condition for the uniqueness of solution of primal and dual problems in (4.2). We recall the following definition from [29]. The set F is said to be *normally smooth* if and only if for every boundary point $\bar{x}$ of F, the *normal cone* of F at $\bar{x}$ defined by $N(\bar{x}; F) := \{u \in \mathbb{R}^n : \ \langle u, x - \bar{x}\rangle \leq 0, \ \forall x \in F\}$ is generated exactly by one vector. That means, there exists $a_{\bar{x}} \in \mathbb{R}^n$ such that $N(\bar{x}; F) = \text{cone}\,\{a_{\bar{x}}\}$. From [29, Proposition 3.3], we have that F is normally smooth if and only if its polar F° is strictly convex.

Proposition 4.2 *Let Q be a nonempty, closed convex subset of $\mathbb{R}^n$. We have the following:*

(i) If Q is strictly convex or if F is strictly convex, then the primal problem in (4.2) has a unique solution, i.e., the generalized projection set $\Pi_F(0; Q)$ is a singleton.
(ii) If F is normally smooth, then the dual problem in (4.2) has a unique solution.

Proof (i) Let $\bar{u}$ be a dual solution in (4.2). Since Q is nonempty and closed, the set $\Pi_F(0; Q)$ is nonempty by Lemma 4.1. Consider the case where Q is strictly convex. Suppose that $\Pi_F(0; Q)$ contains two distinct elements $q_1 \neq q_2$. Then, by relation (4.4), both q_1 and q_2 belong to the set $S_Q(-\bar{u})$. This contradicts Lemma 2.2 by the strictly convexity of Q.
Assume that F is strictly convex. If $0 \in Q$ then $\Pi_F(0; Q) = \{0\}$ is a singleton. Consider the case $0 \notin Q$. Suppose by contradiction that there exist $\bar{q}_1, \bar{q}_2 \in \Pi_F(0; Q)$

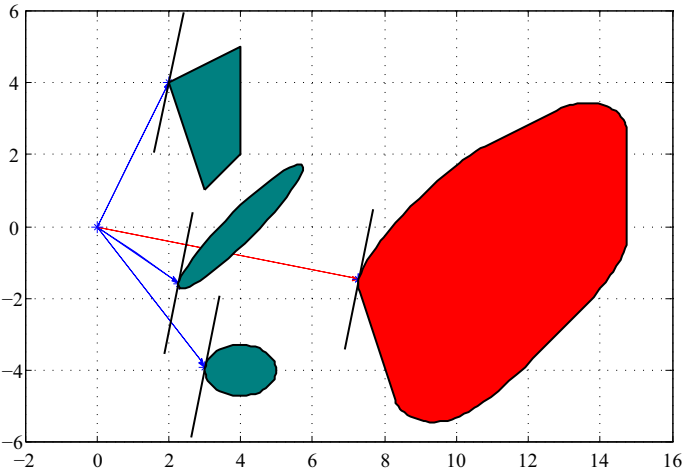


Fig. 3 The Minkowski sum of a polytope and two ellipses is approximately plotted by the largest set. The projection of the origin onto the largest one is the sum of points on three smaller sets such that, at these points, three sets share the same normal vector. This Figure is plotted via Ellipsoidal Toolbox [24]

with $\bar{q}_1 \neq \bar{q}_2$. For $\gamma := T_F(0; Q) > 0$, we have $\rho_F(\bar{q}_1) = \rho_F(\bar{q}_2) = \gamma > 0$. By the positive homogeneity of ρ_F, we have $\rho_F(\frac{\bar{q}_1}{\gamma}) = \rho_F(\frac{\bar{q}_2}{\gamma}) = 1$. This implies $\frac{\bar{q}_1}{\gamma}, \frac{\bar{q}_2}{\gamma} \in F$ and therefore $\frac{1}{2}\left(\frac{\bar{q}_1}{\gamma} + \frac{\bar{q}_2}{\gamma}\right) \in \text{int}(F)$ by the strictly convexity of F. It follows again by the homogeneity of ρ_F that $\rho_F\left(\frac{\bar{q}_1+\bar{q}_2}{2}\right) < \gamma = T_F(0; Q) = \inf\{\rho_F(q) : q \in Q\}$. This is a contradiction. Thus, (i) has been justified. The proof of (ii) is similar by using the strictly convexity of F°. □

From above result, when F is strictly convex and also normally smooth (for example, F is an Euclidean ball or an Ellipsoid), both the primal and dual problems in (4.2) have a unique solution.

Minkowski sum of two closed sets is not necessarily closed. For example, for $Q_1 = \{x \in \mathbb{R}^2 : \ x_2 \geq e^{x_1}\}$ and $Q_2 = \{x \in \mathbb{R}^2 : \ x_2 = 0\}$, the sum

$$Q_1 + Q_2 = \{x \in \mathbb{R}^2 : \ x_2 > 0\}$$

is an open set. In what follows, in order to ensure the existence of support point for the Minkowski sum, we assume that all component sets are compact.

We now show that (4.4) can allow us to characterize points on each constituent sets in the Minkowski sum so that their sum is equal to the projection point; see Fig. 3 for an illustration.

Corollary 4.1 *Let* $\{Q_i\}_{i=1}^p$ *be a finite collection of nonempty convex compact sets in* $\mathbb{R}^n$. *It holds that*

$$T_F\left(0; \sum_{i=1}^p Q_i\right) = -\min\left\{\sum_{i=1}^p \sigma_{Q_i}(-u) : \ u \in F^\circ\right\}.$$

Moreover, if the minimum on the right hand side is attained at $\bar{u} \in F^\circ$*, then any generalized projection* $\bar{q}$ *of the origin onto the set* $\sum_{i=1}^p Q_i$ *satisfies*

$$\bar{q} \in S_{Q_1}(-\bar{u}) + \cdots + S_{Q_p}(-\bar{u}).$$

Thus, the projection $\bar{q}$ *is the sum of points on component sets such that at these points all the sets have the same normal vector* $-\bar{u}$*. If* $F = \mathbb{B}$ *is the Euclidean closed unit ball, then the projection* $\bar{q}$ *exists uniquely and*

$$d\left(0; \sum_{i=1}^p Q_i\right) = -\min\left\{\sum_{i=1}^p \sigma_{Q_i}(-u) : \ u \in \mathbb{B}\right\}.$$

If in addition, $0 \notin \sum_{i=1}^p Q_i$ *then* $\frac{\bar{q}}{\|\bar{q}\|}$ *is the unique solution of the dual problem and we have*

$$\bar{q} \in S_{Q_1}(-\bar{q}) + \cdots + S_{Q_p}(-\bar{q}).$$

Proof Let $Q := \sum_{i=1}^p Q_i$. Using Lemmas 2.1 and 2.2, we have

$$\sigma_Q(-u) = \sigma_{Q_1}(-u) + \cdots + \sigma_{Q_p}(-u) \ \text{ and } \ S_Q(-u) = S_{Q_1}(-u) + \cdots + S_{Q_p}(-u).$$

Note that, the support point mapping $S_Q(u)$ does not depend on the magnitude of u, using Proposition 4.1 and relation (4.4), we clarify the desired conclusion easily. □

The problem of finding a pair of closest points, and therefore the Euclidean distance, between two given convex compact sets $\mathcal{P}$ and $\mathcal{Q}$ can be reduced to the minimum norm problem associated with the Minkowski sum $\mathcal{Q} - \mathcal{P}$ by observing that $d(\mathcal{P}, \mathcal{Q}) = d(0, \mathcal{Q} - \mathcal{P})$. A note here is that although there may be several pairs of closest points, the latter problem always has a unique solution which is the projection from 0 onto $\mathcal{Q} - \mathcal{P}$. By noting that $\sigma_{-\mathcal{P}}(-u) = \sigma_{\mathcal{P}}(u)$ and $S_{-\mathcal{P}}(-u) = -S_{\mathcal{P}}(u)$, we have the following result.

Corollary 4.2 *Let* $\{Q_i\}_{i=1}^p$ *and* $\{P_j\}_{j=1}^\ell$ *be two finite collection of nonempty convex compact sets in* $\mathbb{R}^n$ *and let* $\mathcal{Q} = \sum_{i=1}^p Q_i, \mathcal{P} = \sum_{j=1}^\ell P_j$*. It holds that*

$$d(\mathcal{P}, \mathcal{Q}) = -\min\left\{\sum_{i=1}^p \sigma_{Q_i}(-u) + \sum_{j=1}^\ell \sigma_{P_j}(u) : \ u \in \mathbb{B}\right\}.$$

Moreover, if $\bar{q}$ *is the projection of the origin onto* $Q := \mathcal{Q} - \mathcal{P}$ *and if* $(\bar{a}, \bar{b})$ *is a pair of closest points of* $\mathcal{Q}$ *and* $\mathcal{P}$*, then* $\bar{q} = \bar{a} - \bar{b}$ *and*

$$\bar{a} \in S_{Q_1}(-\bar{q}) + \cdots + S_{Q_p}(-\bar{q}) \quad \text{and} \quad \bar{b} \in S_{P_1}(\bar{q}) + \cdots + S_{P_\ell}(\bar{q}).$$

Thus, $\bar{a}$ is the sum of points in Q_i for $i = 1, \ldots, p$ such that at these points all Q_i have the same normal vector $-\bar{q}$ and $\bar{b}$ is the sum of points in P_j for $j = 1, \ldots, \ell$ such that at these points all P_j have the same normal vector $\bar{q}$.

5 The Gilbert algorithm

We now give an overview and clarify how the Gilbert algorithm can be applied for solving (1.1). Let us define the function $g : \mathbb{R}^n \times Q \to \mathbb{R}$ by

$$g_Q(z, x) := \sigma_Q(z) - \langle z, x\rangle,$$

where $Q = \sum_{i=1}^{p} T_i(\Omega_i)$. From the definition, $g_Q(-z, z) \geq 0$ for all $z \in Q$. A point $z \in Q$ is the solution of (1.1) if and only if $\langle -z, x - z\rangle \leq 0$ for all $x \in Q$. This amounts to saying that $g_Q(-z, z) = 0$.

Lemma 5.1 *If two points z and $\bar{z}$ satisfy $\|z\|^2 - \langle z, \bar{z}\rangle > 0$, then there is a point $\tilde{z}$ in the line segment* conv$\{z, \bar{z}\}$ *such that $\|\tilde{z}\| < \|z\|$.*

Proof If $\|\bar{z}\|^2 \leq \langle z, \bar{z}\rangle$, then we can choose $\tilde{z} = \bar{z}$. Consider the case $\|\bar{z}\|^2 > \langle z, \bar{z}\rangle$. By combining with the assumption $\|z\|^2 - \langle z, \bar{z}\rangle > 0$, we have

$$0 < \lambda^* := \frac{\|z\|^2 - \langle z, \bar{z}\rangle}{\|z - \bar{z}\|^2} < 1.$$

This implies the quadratic function

$$f(\lambda) = \|\bar{z} - z\|^2\lambda^2 + 2\langle z, \bar{z} - z\rangle\lambda + \|z\|^2$$

attains its minimum on $[0, 1]$ at λ^* and therefore $f(\lambda^*) = \|z + \lambda^*(\bar{z} - z)\|^2 < f(0) = \|z\|^2$. Thus $\tilde{z} := z + \lambda^*(\bar{z} - z)$ is the desired point. □

The Gilbert algorithm can be interpreted as follows. Starting from some $z \in Q$, if $g_Q(-z, z) = 0$ then z is the solution. If $g_Q(-z, z) > 0$, then $\bar{z} \in S_Q(-z)$ satisfies $\|z\|^2 - \langle z, \bar{z}\rangle > 0$. Using Lemma 5.1, we find a point $\tilde{z}$ on the line segment connecting z and $\bar{z}$ such that $\|\tilde{z}\| < \|z\|$. The algorithm is outlined as follows.

Gilbert's Algorithm
Initialization: Take an arbitrary point $z_0 \in Q$.
1. If $g_Q(-z, z) = 0$, then return $z = x^*$ is the solution else, set $\bar{z} \in S_Q(-z)$.
2. Compute $\tilde{z} \in$ conv$\{z, \bar{z}\}$ which has minimum norm, set $z = \tilde{z}$ and go back to step 1.

Figure 4 illustrates some iterations of Gilbert's algorithm for finding closest point to an ellipse in two dimension. Lemma 5.1 also suggests an effective way to find $\tilde{z}$ in step 2. We have $\tilde{z} := z + \lambda^*(\bar{z} - z)$, where

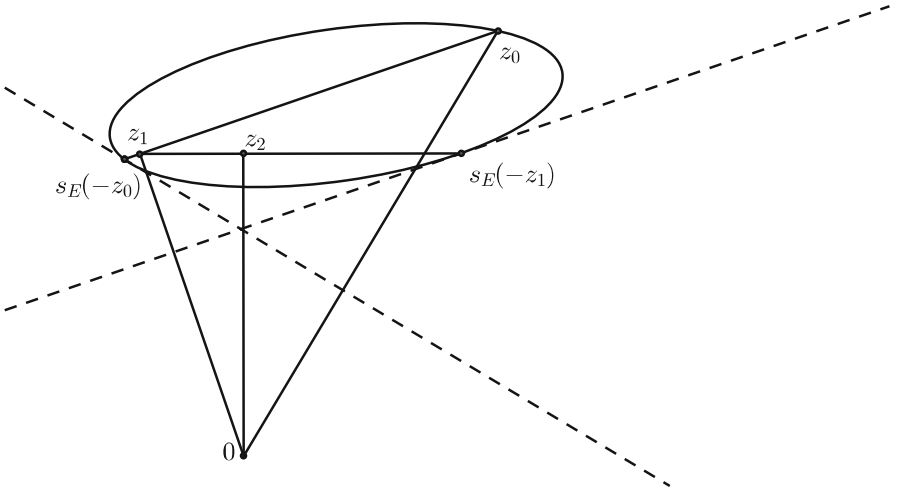


Fig. 4 An illustration of Gilbert's algorithm

$$\lambda^* = \begin{cases} 1, & \text{if } \|\bar{z}\|^2 \leq \langle z, \bar{z}\rangle, \\ \dfrac{\|z\|^2 - \langle z, \bar{z}\rangle}{\|z - \bar{z}\|^2}, & \text{otherwise.} \end{cases} \tag{5.1}$$

To implement the algorithm, it remains to show how to compute a supporting point for $Q = \sum_{i=1}^{p} T_i(\Omega_i)$. Fortunately, this can be done by using Lemma 2.2.
Gilbert showed that, if $\{z_k\}_{k=1}^{\infty}$ generated by the algorithm does not stop with $z = x^*$ at step 1 within a finite number of iterations, then $z_k \to x^*$ asymptotically. According to [17, Theorem 3], we have

$$\|z_k\| - \|x^*\| \leq \frac{C_1}{k} \quad \text{and} \quad \|z_k - x^*\| \leq \frac{C_2}{\sqrt{k}},$$

where C_1 and C_2 are some positive constants. From the above estimates, in order to find an ϵ - approximate solution, i.e., a point z such that $\|z\| - \|x^*\| \leq \epsilon$, we need to perform the algorithm in $O(\frac{1}{\epsilon})$ iterations.

6 Smoothing algorithm for minimum norm problems

Let us first consider functions of the following type

$$\sigma_{A,Q}(u) = \sup\{\langle Au, x\rangle : \ x \in Q\}, \ u \in \mathbb{R}^n,$$

where A is an $m \times n$ matrix and Q is a closed bounded subset of $\mathbb{R}^m$. Observe that $\sigma_{A,Q}(u)$ is the composition of a linear mapping and the support function of Q. As we will see, this function can be approximated by the following function

$$\sigma^{\mu}_{A,Q}(u) = \sup\left\{\langle Au, x\rangle - \frac{\mu}{2}\|x\|^2 : \ x \in Q\right\}, \ u \in \mathbb{R}^n.$$

The following statement is a directly consequence of Theorem 3.1. However, the approximate function as well as its gradient, in this case, has closed form that is expressed in term of the Euclidean projection. This feature makes it reliable from numerical point of view.

Proposition 6.1 *The function $\sigma^{\mu}_{A,Q}$ has the following explicit representation*

$$\sigma^{\mu}_{A,Q}(u) = \frac{\|Au\|^2}{2\mu} - \frac{\mu}{2}\left[d\left(\frac{Au}{\mu}; Q\right)\right]^2$$

and is continuous differentiable on $\mathbb{R}^n$ with its gradient given by $\nabla\sigma^{\mu}_{A,Q}(u) = A^{\top}P_Q\left(\frac{Au}{\mu}\right)$. The gradient $\nabla\sigma^{\mu}_{A,Q}$ is a Lipschitz function with constant $\ell_{\mu} = \frac{1}{\mu}\|A\|^2$. Moreover,

$$\sigma^{\mu}_{A,Q}(u) \le \sigma_{A,Q}(u) \le \sigma^{\mu}_{A,Q}(u) + \frac{\mu}{2}\|Q\|^2 \text{ for all } u \in \mathbb{R}^n, \tag{6.1}$$

where $\|Q\| := \sup\{\|q\| : \ q \in Q\}$.

Proof We have

$$\begin{aligned}
\sigma^{\mu}_{A,Q}(u) &= \sup\left\{\langle Au, x\rangle - \frac{\mu}{2}\|x\|^2 : \ x \in Q\right\}\\
&= \sup\left\{-\frac{\mu}{2}\left(\|x\|^2 - \frac{2}{\mu}\langle Au, x\rangle\right) : \ x \in Q\right\}\\
&= -\frac{\mu}{2}\inf\left\{\left\|x - \frac{Au}{\mu}\right\|^2 - \frac{\|Au\|^2}{\mu^2} : \ x \in Q\right\}\\
&= \frac{\|Au\|^2}{2\mu} - \frac{\mu}{2}\inf\left\{\left\|x - \frac{Au}{\mu}\right\|^2 : \ x \in Q\right\}\\
&= \frac{\|Au\|^2}{2\mu} - \frac{\mu}{2}\left[d\left(\frac{Au}{\mu}; Q\right)\right]^2.
\end{aligned}$$

Since $\psi(x) := [d(x; Q)]^2$ is a differentiable function satisfying $\nabla\psi(x) = 2[x - P_Q(x)]$ for all $x \in \mathbb{R}^m$, we find from the chain rule that

$$\begin{aligned}
\nabla\sigma^{\mu}_{A,Q}(u) &= \frac{1}{\mu}A^{\top}(Au) - \frac{\mu}{2}\left[\frac{2}{\mu}A^{\top}\left(\frac{Au}{\mu} - P_Q\left(\frac{Au}{\mu}\right)\right)\right]\\
&= A^{\top}P_Q\left(\frac{Au}{\mu}\right).
\end{aligned}$$

From the property of the projection mapping onto convex sets and Cauchy–Schwarz inequality, we find, for any $u, v \in \mathbb{R}^n$, that

$$\begin{aligned}
\|\nabla\sigma^{\mu}_{A,Q}(u) - \nabla\sigma^{\mu}_{A,Q}(v)\|^2 &= \left\| A^\top P_Q\left(\frac{Au}{\mu}\right) - A^\top P_Q\left(\frac{Av}{\mu}\right)\right\|^2 \\
&\leq \|A\|^2 \left\| P_Q\left(\frac{Au}{\mu}\right) - P_Q\left(\frac{Av}{\mu}\right)\right\|^2 \\
&\leq \|A\|^2 \left\langle \frac{Au - Av}{\mu}, P_Q\left(\frac{Au}{\mu}\right) - P_Q\left(\frac{Av}{\mu}\right)\right\rangle \\
&= \frac{\|A\|^2}{\mu}\left\langle u - v, A^\top P_Q\left(\frac{Au}{\mu}\right) - A^\top P_Q\left(\frac{Av}{\mu}\right)\right\rangle \\
&= \frac{\|A\|^2}{\mu}\langle u - v, \nabla\sigma^{\mu}_{A,Q}(u) - \nabla\sigma^{\mu}_{A,Q}(v)\rangle \\
&\leq \frac{\|A\|^2}{\mu}\|u - v\|\|\nabla\sigma^{\mu}_{A,Q}(u) - \nabla\sigma^{\mu}_{A,Q}(v)\|.
\end{aligned}$$

This implies that

$$\|\nabla\sigma^{\mu}_{A,Q}(u) - \nabla\sigma^{\mu}_{A,Q}(v)\| \leq \frac{\|A\|^2}{\mu}\|u - v\|.$$

The lower and upper bounds in (6.1) follow from

$$\langle Au, x\rangle - \frac{\mu}{2}\|x\|^2 \leq \langle Au, x\rangle \leq \langle Au, x\rangle - \frac{\mu}{2}\|x\|^2 + \frac{\mu}{2}\sup\left\{\|q\|^2 : \ q \in Q\right\},$$

for all $x \in Q$. The proof is now complete. □

From Proposition 4.1, we have the duality result

$$d\left(0; \sum_{i=1}^{p} T_i(\Omega_i)\right) = -\min\left\{\sum_{i=1}^{p}\sigma_{\Omega_i}(-A_i^\top u) - \left\langle u, \sum_{i=1}^{p} a_i\right\rangle : \ u \in \mathbb{B}\right\}.$$

This dual problem is a non-smooth convex problem with constraint. This makes it not favorable to apply optimization scheme. We now make use of the strong convexity of the squared Euclidean norm to state another dual problem for (1.1) in which the dual objective function is strongly convex and the constraint is removed.

Proposition 6.2 *The following duality result holds*

$$\left[d\left(0; \sum_{i=1}^{p} T_i(\Omega_i)\right)\right]^2 = -\min\left\{\sum_{i=1}^{p}\sigma_{\Omega_i}\left(A_i{}^\top u\right) + \left\langle u, \sum_{i=1}^{p} a_i\right\rangle + \frac{1}{4}\|u\|^2 : \ u \in \mathbb{R}^n\right\}. \tag{6.2}$$

Moreover if u^* is the unique solution of the dual problem in (6.2), then $x^* = -\frac{1}{2}u^*$.

Proof By observing that the dual of the function $\|\cdot\|^2$ is $\frac{1}{4}\|\cdot\|^2$, applying Theorem 2.1 for $Q := \sum_{i=1}^{p} T_i(\Omega_i)$, we have

$$\begin{aligned}
[d\,(0;\,Q)]^2 &= \inf\left\{\|x\|^2 :\ x \in Q\right\} \\
&= \max\left\{-\left(\delta_Q\right)^*(u) - \left(\|\cdot\|^2\right)^*(-u) :\ u \in \mathbb{R}^n\right\} \\
&= \max\left\{-\sigma_Q(u) - \frac{1}{4}\|-u\|^2 :\ u \in \mathbb{R}^n\right\} \\
&= -\min\left\{\sigma_Q(u) + \frac{1}{4}\|u\|^2 :\ u \in \mathbb{R}^n\right\}.
\end{aligned}$$

Equality (6.2) now follows directly from Lemma 2.1. The objective dual function in (6.2) is strongly convex, solution u^* exists uniquely. Recall that $x^* = P_Q(0)$, from this duality result, we have $\|x^*\|^2 = -\sigma_Q(u^*) - \frac{1}{4}\|u^*\|^2$. Moreover, $\sigma_Q(u^*) = \sup_{q\in Q}\langle u^*, q\rangle \geq \langle u^*, x^*\rangle$ since $x^* \in Q$. We have $\|x^*\|^2 + \frac{1}{4}\|u^*\|^2 \leq -\langle u^*, x^*\rangle$ which is equivalent to $\|x^* + \frac{1}{2}u^*\|^2 \leq 0$ and therefore $x^* = -\frac{1}{2}u^*$. □

In order to solve minimum norm problem (1.1), we solve dual problem (6.2) by approximating the dual objective function by a smooth and strongly convex function with Lipschitz continuous gradient and then apply a fast gradient scheme to this smooth one.
Let us define the dual objective function by

$$f(u) := \sum_{i=1}^{p} \sigma_{\Omega_i}\left(A_i{}^{\top}u\right) + \langle u, \sum_{i=1}^{p} a_i\rangle + \frac{1}{4}\|u\|^2, \quad u \in \mathbb{R}^n.$$

The following result is a direct consequence of Proposition 6.1.

Proposition 6.3 *The function $f(u)$ has the following smooth approximation*

$$f_\mu(u) := \sum_{i=1}^{p}\left(\frac{\|A_i^{\top}u\|^2}{2\mu} - \frac{\mu}{2}\left[d(\frac{A_i^{\top}u}{\mu};\,\Omega_i)\right]^2\right) + \left\langle u, \sum_{i=1}^{p} a_i\right\rangle + \frac{1}{4}\|u\|^2, \quad u \in \mathbb{R}^n.$$

Moreover, f_μ is a strongly convex function with modulus $\gamma = \frac{1}{2}$ and its gradient is given by

$$\nabla f_\mu(u) = \sum_{i=1}^{p} A_i\, P_{\Omega_i}\left(\frac{A_i^{\top}u}{\mu}\right) + \sum_{i=1}^{p} a_i + \frac{1}{2}u. \tag{6.3}$$

The Lipschitz constant of ∇f_μ is

$$L_\mu := \frac{\sum_{i=1}^{p}\|A_i\|^2}{\mu} + \frac{1}{2}. \tag{6.4}$$

Moreover, we have the following estimate

$$f_\mu(u) \le f(u) \le f_\mu(u) + \mu D_f,$$

where $D_f := \frac{1}{2}\sum_{i=1}^{p} \|\Omega_i\|^2 < \infty$.

We now apply the Nesterov fast gradient method introduced in Sect. 3 to minimize f_μ. The **NE**sterov **S**moothing algorithm for **MI**nimum **NO**rm problem (NESMINO) is outlined as follows:

NESMINO
INITIALIZE: Ω_i, A_i, a_i for $i = 1, \ldots, p$ and v_0, u_0, μ.
Set $k = 0$.
Repeat the following
Compute $\nabla f_\mu(v_k)$ using (6.3)
Compute L_μ using (6.4)
Set $u_{k+1} := v_k - \frac{1}{L_\mu}\nabla f_\mu(v_k)$
Set $v_{k+1} := u_{k+1} + \frac{\sqrt{L_\mu}-\sqrt{1/2}}{\sqrt{L_\mu}+\sqrt{1/2}}(u_{k+1} - u_k)$
Set $k := k + 1$
Until a stopping criterion is satisfied.

We denote by u_μ^* the unique minimizer of f_μ on $\mathbb{R}^n$. We also denote by u^* a minimizer of f and by $f^* := f(u^*) = \inf_{x\in\mathbb{R}^n} f(x)$ its optimal value on $\mathbb{R}^n$. From the duality result (6.2), we have

$$f^* = -\left[d\left(0; \sum_{i=1}^{p} T_i(\Omega_i)\right)\right]^2 = -\|x^*\|^2. \tag{6.5}$$

Recall that, our objective function in primal problem (1.1) is the Euclidean norm function $\|\cdot\|$. A feasible point $x \in \sum_{i=1}^{p} T_i(\Omega_i)$ is said to be an ϵ-approximate solution of problem (1.1) if it satisfies

$$\|x\| - \|x^*\| \le \epsilon.$$

Our purpose is to solve the original minimum norm problem (1.1) with an accuracy ϵ. From the structure of (1.1), it is very challenging for us to deal with the constraint $x \in \sum_{i=1}^{p} T_i(\Omega_i)$ and therefore we have employed the dual approach. In our approach, the one that we are optimizing is the smooth approximation of the dual objective function. It remains to show how to recover an approximate primal solution from the dual iterative sequence.

Theorem 6.1 *Let* $\{u_k\}_{k=1}^{\infty}$ *be the sequence generated by* NESMINO *algorithm. Then the sequence* $\{x_k\}_{k=1}^{\infty}$ *defined by*

$$x_k := \sum_{i=1}^{p}\left[A_i P_{\Omega_i}\left(\frac{A_i^\top u_k}{\mu}\right) + a_i\right]$$

converges to an ϵ-approximate solution of minimum norm problem (1.1) within $k = O\left(\frac{1}{\sqrt{\epsilon}} \ln\left(\frac{1}{\epsilon}\right)\right)$ iterations.

Proof It follows from [31, Theorem 2.2.3] that the iterative sequence $\{u_k\}_{k=0}^{\infty}$ satisfies

$$f_\mu(u_k) - f_\mu^* \leq 2\left(f_\mu(u_0) - f_\mu^*\right) e^{-k\sqrt{\frac{\gamma}{L_\mu}}}. \tag{6.6}$$

From $f_\mu(u_0) \leq f(u_0)$ and the following estimate

$$f_\mu^* = f_\mu(u_\mu^*) \geq f(u_\mu^*) - \mu D_f \geq f(u^*) - \mu D_f = f^* - \mu D_f,$$

we have

$$f_\mu(u_0) - f_\mu^* \leq f(u_0) - f^* + \mu D_f. \tag{6.7}$$

Moreover, since $f_\mu(u_k) - f_\mu^* \geq f(u_k) - \mu D_f - f^*$, we find from (6.6) and (6.7) that

$$\begin{aligned} f(u_k) - f^* &\leq \mu D_f + f_\mu(u_k) - f_\mu^* \\ &\leq \mu D_f + 2\left(f(u_0) - f^* + \mu D_f\right) e^{-k\sqrt{\frac{\gamma}{L_\mu}}}, \text{ for all } k \geq 0. \end{aligned} \tag{6.8}$$

Since f_μ is a differentiable strongly convex function and u_μ^* is its unique minimizer on $\mathbb{R}^n$, we have $\nabla f_\mu\left(u_\mu^*\right) = 0$. It follows from [31, Theorem 2.1.5] that

$$\frac{1}{2L_\mu}\|\nabla f_\mu(u_k)\|^2 \leq f_\mu(u_k) - f_\mu^* \overset{(6.6)}{\leq} 2\left(f_\mu(u_0) - f_\mu^*\right) e^{-k\sqrt{\frac{\gamma}{L_\mu}}}.$$

This implies

$$\begin{aligned} \|\nabla f_\mu(u_k)\|^2 &\leq 4L_\mu(f_\mu(u_0) - f_\mu^*)e^{-k\sqrt{\frac{\gamma}{L_\mu}}} \\ &\overset{(6.7)}{\leq} 4L_\mu(f(u_0) - f^* + \mu D_f)e^{-k\sqrt{\frac{\gamma}{L_\mu}}}. \end{aligned} \tag{6.9}$$

For each k and for each $i \in \{1, \ldots, p\}$, let w_k^i be the unique solution to the problem

$$\sigma_{\mu,\Omega_i}\left(A_i^\top u_k\right) := \sup\left\{\langle A_i^\top u_k, w\rangle - \frac{\mu}{2}\|w\|^2 : w \in \Omega_i\right\}.$$

We have

$$\begin{aligned}
&\sup\left\{\langle A_i^\top u_k, w\rangle - \frac{\mu}{2}\|w\|^2 : w \in \Omega_i\right\} \\
&= \sup\left\{\frac{\|A_i^\top u_k\|^2}{2\mu} - \frac{\mu}{2}\left\|\frac{A_i^\top u_k}{\mu} - w\right\|^2 : w \in \Omega_i\right\} \\
&= \frac{\|A_i^\top u_k\|^2}{2\mu} - \frac{\mu}{2}\left[d\left(\frac{A_i^\top u_k}{\mu}, \Omega_i\right)\right]^2.
\end{aligned}$$

Hence $w_k^i = P_{\Omega_i}\left(\frac{A_i^\top u_k}{\mu}\right)$. For each k, we have

$$d_k := \left\|\sum_{i=1}^p \left(A_i w_k^i + a_i\right)\right\|^2 - \left[d\left(0, \sum_{i=1}^p T_i(\Omega_i)\right)\right]^2 = \left\|\sum_{i=1}^p \left(A_i w_k^i + a_i\right)\right\|^2 + f^*,$$

where the equality is due to (6.5). Observe that the sequence $\{x_k\}$ is primal feasible, i.e., $x_k = \sum_{i=1}^p \left(A_i w_k^i + a_i\right) \in \sum_{i=1}^p T_i(\Omega_i)$. From the property of the projection onto convex sets, we have $\langle -x^*, x_k - x^*\rangle \le 0$ and hence

$$\|x_k - x^*\|^2 = \|x_k\|^2 - \|x^*\|^2 + 2\left\langle -x^*, x_k - x^*\right\rangle \le \|x_k\|^2 - \|x^*\|^2 = d_k.$$

This implies that $\{x_k\}$ converges to x^* whenever $d_k \to 0$ as $k \to \infty$. Moreover, we have

$$2\|x^*\|\left(\|x_k\| - \|x^*\|\right) \le \left(\|x_k\| + \|x^*\|\right)\left(\|x_k\| - \|x^*\|\right) = \|x_k\|^2 - \|x^*\|^2 = d_k. \tag{6.10}$$

We have the following

$$\begin{aligned}
d_k &= \left\|\sum_{i=1}^p (A_i w_k^i + a_i)\right\|^2 + f^* \\
&= \left\|\sum_{i=1}^p \left(A_i w_k^i + a_i\right)\right\|^2 + f_\mu(u_k) + f^* - f_\mu(u_k) \\
&= \left\|\sum_{i=1}^p (A_i w_k^i + a_i)\right\|^2 + \sum_{i=1}^p \left[\left\langle A_i^\top u_k, w_k^i\right\rangle - \frac{\mu}{2}\|w_k^i\|^2\right] \\
&\quad + \left\langle u_k, \sum_{i=1}^p a_i\right\rangle + \frac{1}{4}\|u_k\|^2 + f^* - f_\mu(u_k)
\end{aligned}$$

$$
\begin{aligned}
&= \left\| \sum_{i=1}^{p} (A_i w_k^i + a_i) \right\|^2 + \left\langle u_k, \sum_{i=1}^{p} (A_i w_k^i + a_i) \right\rangle + \frac{1}{4}\|u_k\|^2 - \frac{\mu}{2} \sum_{i=1}^{p} \|w_k^i\|^2 \\
&\quad + f^* - f_\mu(u_k) \\
&= \left\| \sum_{i=1}^{p} \left(A_i w_k^i + a_i \right) + \frac{1}{2} u_k \right\|^2 - \frac{\mu}{2} \sum_{i=1}^{p} \|w_k^i\|^2 + f^* - f_\mu(u_k) \\
&= \left\| \sum_{i=1}^{p} A_i P_{\Omega_i}\left(\frac{A_i^\top u_k}{\mu}\right) + \sum_{i=1}^{p} a_i + \frac{1}{2} u_k \right\|^2 - \frac{\mu}{2} \sum_{i=1}^{p} \|w_k^i\|^2 + f^* - f_\mu(u_k) \\
&= \|\nabla f_\mu(u_k)\|^2 - \frac{\mu}{2} \sum_{i=1}^{p} \|w_k^i\|^2 + f^* - f_\mu(u_k).
\end{aligned}
$$

Observe that $|f_\mu(u_k) - f^*| \overset{(6.3)}{\leq} |f(u_k) - f^*| + \mu D_f$ and $\sum_{i=1}^{p} \|w_k^i\|^2 \leq 2D_f$, taking into account (6.8) and (6.9), we have

$$
\begin{aligned}
d_k &\leq \left\|\nabla f_\mu(u_k)\right\|^2 + |f(u_k) - f^*| + 2\mu D_f \\
&\leq 4L_\mu (f(u_0) - f^* + \mu D_f) e^{-k\sqrt{\frac{\gamma}{L_\mu}}} + \mu D_f \\
&\quad + 2(f(u_0) - f^* + \mu D_f) e^{-k\sqrt{\frac{\gamma}{L_\mu}}} + 2\mu D_f. \\
&\leq 2(2L_\mu + 1)\left(f(u_0) - f^* + \mu D_f\right) e^{-k\sqrt{\frac{\gamma}{L_\mu}}} + 3\mu D_f.
\end{aligned}
$$

For a fix $\epsilon > 0$, from (6.10) in order to achieve an ϵ - approximate solution for the primal problem, we should force each of the two terms in the above estimate less than or equal to $\frac{\epsilon}{2}$. If we choose the value of smooth parameter μ to be $\frac{\epsilon}{6D_f}$, we have $d_k \leq \epsilon$ whenever

$$
k \geq \sqrt{\frac{L_\mu}{\gamma}} \ln\left(\frac{4(2L_\mu + 1)\left(f(u_0) - f^* + \frac{\epsilon}{6}\right)}{\epsilon} \right), \tag{6.11}
$$

where $L_\mu = \frac{\sum_{i=1}^{p} 6\|A_i\|^2 D_f}{\epsilon} + \frac{1}{2}$. Thus, we can find an ϵ - approximate solution for primal problem within $k = O\left(\frac{1}{\sqrt{\epsilon}} \ln\left(\frac{1}{\epsilon}\right)\right)$ iterations. The proof is complete. □

Remark 6.1 In NESMINO algorithm, a smaller smooth parameter μ is often better because it reduces the error when approximate f by f_μ. However, a small μ implies a large value of the Lipschitz constant L_μ which in turn reduces the convergence rate by (6.11). Thus the time cost of the algorithm is expensive if we fix a value for μ ahead of time. In practice, a sequence of smooth problems with decreasing smooth parameter μ is solved and the solution of the previous problem is used as the initial point for the next one. The algorithm stops when a preferred μ_* is attained. The optimization scheme is outlined as follows.

INITIALIZE: Ω_i, A_i, a_i for $i = 1, \ldots, p$ and $w_0, \sigma \in (0, 1)$, $\mu_0 > 0$ and $\mu_* > 0$. Set $k = 0$.
Repeat the following
1. Apply NESMINO algorithm with $\mu = \mu_k$, $u_0 = v_0 = w_k$ to find $w_{k+1} = \operatorname{argmin}_{w \in \mathbb{R}^n} f_\mu(w)$.
2. Update $\mu_{k+1} := \sigma \mu_k$ and set $k := k + 1$.

Until $\mu \leq \mu_*$.

We highlight the fact that the algorithm does not require computation of the Minkowski sum but rather only the projection onto each of the constituent sets Ω_i. Fortunately, many useful projection operators are easy to compute. Explicit formula for projection operator P_Ω exists when Ω is a closed Euclidean ball, a closed rectangle, a hyperplane, or a half-space. Although there are no analytic solutions, fast algorithms for computing the projection operators exist for the cases of unit simplex, the closed ℓ_1 ball (see [8,12]), or the ellipsoids (see [9]).

An advantage of smoothing approach is that in many cases, by making use of the special structure of the support function of Ω, we can have a suitable smoothing technique that can avoid working with implicit projection operator P_Ω or employ some fast projection algorithm. We consider two important cases as follows:
The case of ellipsoids Consider the case of ellipsoids associated with Euclidean norm

$$E(A, c) := \left\{ x \in \mathbb{R}^n : \ (x - c)^\top A^{-1}(x - c) \leq 1 \right\},$$

where the shape matrix A is positive definite and the center c is some given point in $\mathbb{R}^n$. It is well known that the support function of this Ellipsoid is $\sigma_E(u) = \sqrt{u^\top A u} + u^\top c$ and the support point in direction u is $\frac{Au}{\sqrt{u^\top A u}} + c$. We can rewrite the support function as follows

$$\sigma_E(u) = \|A^{1/2}u\| + u^\top c = \sigma_{\mathbb{B}}\left(A^{1/2}u\right) + u^\top c,$$

where $\mathbb{B}$ stands for the closed unit Euclidean ball and $A^{1/2}$ is the square root of A. The smooth approximation g_μ of the function $g = \sigma_E$ has the following explicit representation

$$g_\mu(u) = \frac{\|A^{1/2}u\|^2}{2\mu} - \frac{\mu}{2}\left[d(\frac{A^{1/2}u}{\mu}; \mathbb{B})\right]^2 + u^\top c.$$

and is differentiable on $\mathbb{R}^n$ with its gradient given by $\nabla g(u) = A^{1/2} P_{\mathbb{B}}\left(\frac{A^{1/2}u}{\mu}\right) + c$. Thus, instead of projecting onto the Ellipsoid, we just need to project onto the closed unit ball.
The case of polytopes Consider the polytope $S = \operatorname{conv}\{a_1, \ldots, a_m\}$ generated by m point in $\mathbb{R}^n$. By [34, Theorem 32.2], we have

$$\sigma_S(u) = \sup\{\langle u, x\rangle : \ x \in S\} = \max_{1 \leq i \leq m} \langle u, a_i\rangle,$$

and the support point S is some point a_i such that $\langle u, a_i\rangle = \sigma_S(u)$. For $\alpha = (\alpha_1, \ldots, \alpha_m)^\top \in \mathbb{R}^m$, we have

$$\max_{1\le i\le m} \alpha_i = \sup\left\{x_1\alpha_1 + \cdots + x_m\alpha_m : x_i \ge 0, \sum_{i=1}^m x_i = 1\right\} = \sup\{\langle \alpha, x\rangle : x \in \Delta_m\}.$$

Therefore, $\sigma_S(u) = \sup\{\langle Au, x\rangle : x \in \Delta_m\} = \sigma_{\Delta_m}(Au)$, where $A = [a_1, a_2, \ldots, a_m]^\top$ is an $m \times n$ matrix whose i^{th} row is $a_i^\top$ and Δ_m is the unit simplex in $\mathbb{R}^m$. The smooth approximate function of $g = \sigma_S$ is $g_\mu(u) = \dfrac{\|Au\|^2}{2\mu} - \dfrac{\mu}{2}\left[d(\dfrac{Au}{\mu}; \Delta_m)\right]^2$, with $\nabla g_\mu(u) = A^\top P_{\Delta_m}\left(\dfrac{Au}{\mu}\right)$. We thus can employ the fast and simple algorithms for computing the projection onto a unit simplex, for example in [7,8], instead of projection onto a polytope.

Remark 6.2 The classical Frank-Wolfe method for solving the problem $\min\{f(x) : x \in Q\}$ has the following form

$$x_{k+1} = x_k + \lambda_k(s_k - x_k),$$

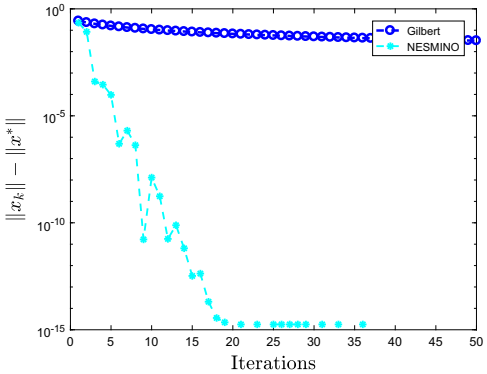
where $s_k = \operatorname{argmin}\{\langle \nabla f(x_k), s\rangle : s \in Q\}$; see [13,21]. Gilbert's algorithm can be seen as a Frank-Wolfe type method in which $f(x) = \frac{1}{2}\|x\|^2$ and the step-size sequence λ_k is chosen in a special way according to (5.1). It is well-known that the FW method has slow convergence rate of $O(1/k)$ because of the so-called zig-zagging phenomenon [21,22]. Especially when the optimal solution x^* does not lie in the relative interior of Q, the FW iterate tends to zig-zag amongst the vertices that define the face containing x^*. Theorem 6.1 shows that, the complexity bound $O\left(\frac{1}{\sqrt{\epsilon}}\ln(\frac{1}{\epsilon})\right)$ of NESMINO is better than the worst-case complexity $O(\frac{1}{\epsilon})$ of Gilbert's algorithm. Moreover, we can also reduce the expensive task of projection onto original constituent sets in NESMINO to projection onto some much simpler ones. Very recently, Garber and Hazan [16] proved that in case where the feasible set Q is strongly convex, the Frank-Wolfe method converges at an accelerated rate of $O(\frac{1}{k^2})$. The step-size λ_k in [16, Algorithm 1] is updated by the following rule

$$\lambda_k \longleftarrow \operatorname*{argmin}_{\lambda\in[0,1]} \lambda\langle s_k - x_k, \nabla f(x_k)\rangle + \lambda^2\frac{\beta_f}{2}\|s_k - x_k\|^2.$$

From the proof of Lemma 5.1, this is exactly the same as Gilbert's algorithm with $f(x) = \frac{1}{2}\|x\|^2$ and $\beta_f = 1$.

The rest of this section is devoted to conducting some numerical examples for both NESMINO and Gilber's algorithm. All the tests are implemented by MATLAB R2016b on a personal computer with an Intel Core i5 CPU 1.6 GHz and 4G of RAM.

Example 6.1 Let us first consider a simple example where the optimal solution is known in advance. Consider the minimum norm problem associated with a polytope



Iter.	NESMINO	Gilbert
1	(1.5, 1.5)	(1.5, 1.5)
10	(0, 1)	(0.0611, 1.1071)
100	(0, 1)	(0.0090, 1.0177)
300	(0, 1)	(0.0032, 1.0063)
1000	(0, 1)	(0.0010, 1.0020)
10000	(0, 1)	(0.0001, 1.0002)

Fig. 5 A comparison of NESMINO and Gilbert's algorithm for finding the projection onto a polytope

P in $\mathbb{R}^2$ whose vertices are $(-2, 1)$, $(2, 1)$ and $(1, 2)$. The projection of the origin onto P is $x^* = (0, 1)$. Starting from $(\frac{3}{2}, \frac{3}{2})$, we implement NESMINO and Gilbert's algorithm in 10^4 iterations and report the result in Fig. 5. The NESMINO algorithm, with a fixed value $\mu = 0.1$ converges to the optimal solution x^* within 10 steps. In contrast, the approximate values in Gilbert's algorithm are still changing after 10^4 iterations. In this case, as the number of iterations is increasing, the Gilbert algorithm alternately chooses the two vertices $(-2, 1)$ and $(2, 1)$ as support points of P and turns to be very slow when it approaches the solution x^*.

The above simple case can be modified to get a polytope in arbitrary n dimension to which the Gilbert algorithm usually has the zig-zagging phenomenon. Let $m = 2\ell+1$. We generate a polytope P of m vertices according to the MATLAB syntax: $A = rand(\ell, n-1)$, $B = [[A; -A], ones(2\ell, 1)]$ and $P = [B; [rand(1, n-1), 10]]$. Polytope P in this case has all vertices belonging to the hyperplane $x_n = 1$ except the last one belonging to the hyperplane $x_n = 10$ and the projection of the origin in this case is $x^* = [zeros(n-1, 1); 1]$. For this polytope, NESMINO converges in several steps while Gilbert's algorithm converges very slowly as above.

Example 6.2 We now consider the problem of computing the projection onto a sum of polytopes in high dimension. We generate p polytopes in n dimension space, each of them has m vertices. The vertices of the i^{th} polytope are rows of an $m \times n$ matrix A_i that is randomly generated by the MATLAB function (2*i)*rand(m,n).

For general problem, the exact solution x^* to (1.1) cannot be computed analytically. Evaluating the complexity of the algorithms based on the number of iterations requires defining a nearly optimal solution. For each problem, we first run Gilbert's algorithm in a large enough number of iterations to find such a referenced solution.

(a) Let us first consider the case where $p = 2$. For each value of the pair (m, n), we generated 100 different problems and track the progress of each algorithm during the iteration by computing the relative error $\frac{\|x_k\|-\|x^*\|}{\|x^*\|}$. We implement NESMINO with the geometrically decreasing sequence $\mu_k = 10\left(\frac{1}{2}\right)^k$ of smooth parameter μ, i.e., $\mu_0 = 10$ and $\sigma = \frac{1}{2}$ in Remark 6.1. We switch to the next smaller μ whenever

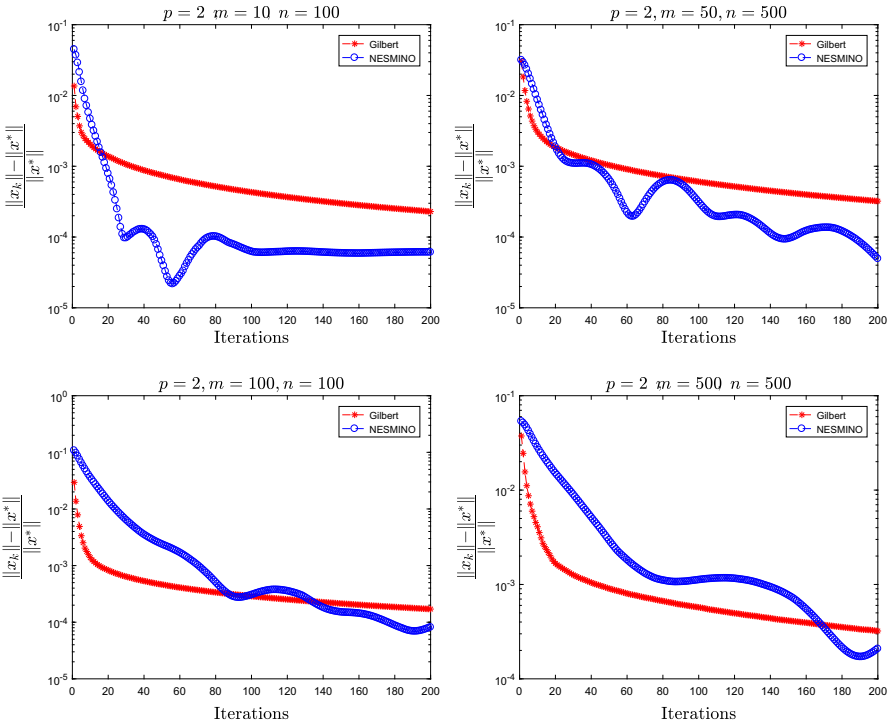


Fig. 6 Performance of NESMINO and Gilbert's algorithm in finding the projection onto the sum of two polytopes. We use $\mu_0 = 10, \sigma = 0.5$ for all cases

$\|\nabla f_\mu(u_k)\| \leq \epsilon = 10^{-3}$. Each step of NESMINO requires to compute the full gradient

$$\nabla f_\mu(v) = \sum_{i=1}^{p} A_i^\top P_{\Delta_m}\left(\frac{A_i v}{\mu}\right) + \frac{1}{2}v.$$

We can reduce computation time without using any for loop by the following line of MATLAB code

P′ ∗ reshape(simplexproj(reshape((1/mu) ∗ P ∗ v,[],p)),[],1)+ (1/2) ∗ v;

where $P = [A_1^\top, \ldots, A_p^\top]^\top \in \mathbb{R}^{pm \times n}$ and simplexproj(Y) is a fast procedure to projection each column in Y onto the unit simplex, see [8].

Figure 6 plots the average of the relative error on a log scale at each iteration of the two methods. From this test, we can see that Gilbert's algorithm usually decreases extremely fast at beginning iterations and turns to be slow after that. In contrast, smoothing algorithm decreases slowly at starting iterations and its improvement will be faster than that of Gilbert's algorithm from a certain iteration k_0. The value of k_0 is larger when the number of vertices m is larger. When m is quite small, the NESMINO

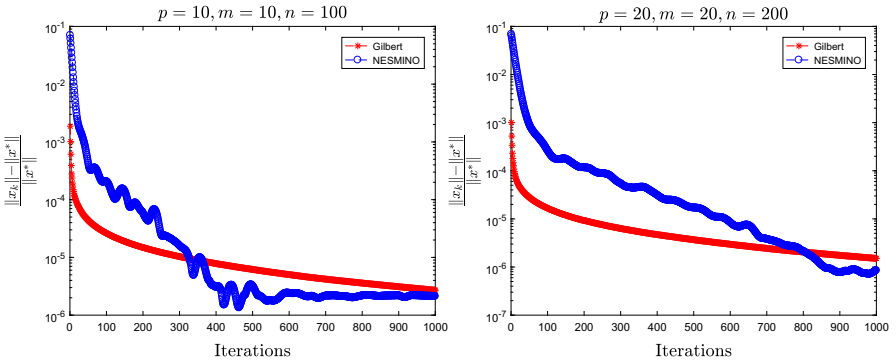


Fig. 7 Performance of NESMINO and Gilbert's algorithm for medium scale problems. We use $\mu_0 = 20, \sigma = 0.5$ for the left and and $\mu_0 = 200, \sigma = 0.5$ for the right

scales well with the space dimensions and performs better than Gilbert's algorithm. NESMINO is quite sensitive to the choice of the smooth parameter μ and therefore to both μ_0 and σ. When the size of the problems is large, increasing the value of μ_0 can significantly improve the performance of NESMINO. Figure 7 shows the results for the cases of larger p and m. We can see that NESMINO performs well at final iterations.

(b) Now we consider the case with very large number of components p in the Minkowski sum. NESMINO gets stuck because its computational cost in each of the iteration turns to be very expensive when p large. An advantage of smoothing approach is that it allows us to use some stochastic methods in this case. For our minimum norm problem with large p, we apply the SAGA [11] to minimize the objective function f_μ instead of fast gradient method as before. The resulting algorithm is called SAGA-NESMINO.

It is known that SAGA is inspired from SAG (Stochastic Average Gradient) [35]. However, instead of using a biased gradient estimate as in SAG, SAGA use an unbiased update direction. Given x_0 and $y_0^i = \nabla f_i(x_0)$ for $i = 1, \ldots, p$, to minimize $f(x) = \frac{1}{p}\sum_{i=1}^{p} f_i(x)$, at the k^{th} iteration, SAGA picks an index j uniformly at random from $\{1, \ldots, p\}$, sets $y_k^i = \nabla f_j(x_{k-1})$ if $i = j$ and $y_k^i = y_{k-1}^i$ otherwise and then updates

$$x_k = x_{k-1} - \alpha \left[y_k^j - y_{k-1}^j + \frac{1}{p}\sum_{i=1}^{p} y_{k-1}^i \right].$$

Recall that our objective function for the case of p polytopes can be written as

$$f_\mu(u) = \sum_{i=1}^{p} f_i(u) = \sum_{i=1}^{p} \left[\frac{\|A_i u\|^2}{2\mu} - \frac{\mu}{2}\left[d(\frac{A_i u}{\mu}; \Delta_m) \right]^2 + \frac{1}{4p}\|u\|^2 \right],$$

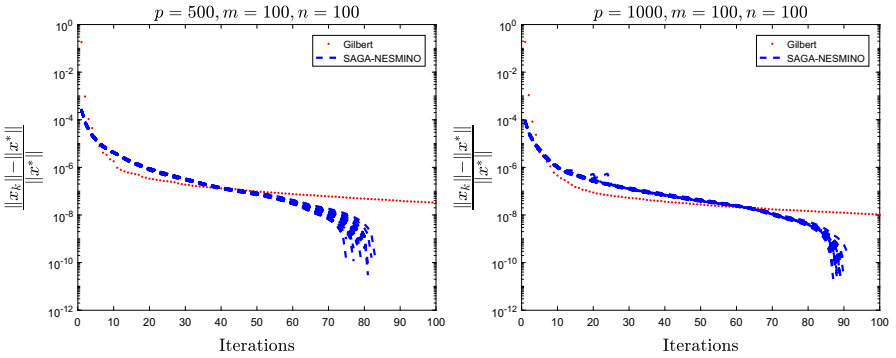


Fig. 8 Comparison between SAGA-NESMINO and Gilbert's algorithm for large scale problems

and the gradient of the i^{th} component is $\nabla f_i(u) = A_i^\top P_{\Delta_m}\left(\frac{A_i u}{\mu}\right) + \frac{1}{2p}u$ for $i = 1, \ldots, p$. Each component function f_i is μ-strongly convex with L-Lipschitz continuous gradient, where $\mu = \frac{1}{2p}$ and $L = \frac{1}{\mu} \max_{i=1,\ldots,p} \|A_i\|^2 + \frac{1}{2p}$. At each iteration, SAGA-NESMINO requires to compute only one projection instead of p projection as in NESMINO. Therefore, the computational cost of SAGA-NESMINO is much cheaper when p is large.

In this test, we set the smooth parameter $\mu = 0.01$, take $x_0 = 0$ and $y_o^i = \nabla f_i(x_0)$ for $i = 1, \ldots, p$ and use the constant step-size $\alpha = \frac{10^4}{64L}$. To have a fair comparison between a stochastic method and a full gradient method, we identify p iterations of SAGA-NESMINO with 1 iteration of Gilbert's algorithm. For each case of (p, m, n), we run SAGA-NESMINO in 20 times and plot the relative error at each iteration on Fig. 8. Once again, we can see that smoothing algorithm can reach a high accuracy approximation solution (for example, with relative error up to 10^{-10}) faster than Gilbert's algorithm.

7 Conclusions

Minimum norm problems have been studied from both theoretical and numerical point of view in this paper. Based on duality approach, it is shown that projection onto a Minkowski sum of sets can be represented as the sum of points on constituent sets, so that at these points, all of the sets share the same normal vector. By combining Nesterov's smoothing technique and his fast gradient scheme, we have developed a numerical algorithm for solving the problems. The proposed NESMINO is proved to have a better complexity bound than the worst-case complexity bound of Gilbert's algorithm. Gilbert's algorithm usually decreases slowly as it approaches the solution and therefore it is slow in finding an approximate solution with high accuracy. In such situations, smoothing-based methods can be seen as a good alternative.

Very recently, Won et al. [40] proposed a simple but efficient block descent algorithm for projecting onto Minkowski sums of sets. The algorithm showed its

competitive performances on solving several statistical learning problems which are instances of (1.2). It is worth noting that, our NESMINO algorithm can allow us to deal with affine transformations on the sets and to avoid working with complex projection operators in many cases. The smoothing technique derived in this paper can also be employed to obtain fast schemes for solving (1.2), especially when the fidelity loss f is differentiable. It is of interest to conduct a comparison between NESMINO and the block descent algorithm in solving (1.2). These are interesting topics for our future research.

Acknowledgements This article was supported by the National Natural Science Foundation of China under Grant Nos. 11401152 and 11950410503. Research of the second author was supported by the China Postdoctoral Science Foundation under Grant No. 2017M622991 and the Vietnam National Foundation for Science and Technology Development under Grant No. 101.01-2017.325. The authors would like to thank anonymous reviewers for insightful comments that helped to greatly improve the manuscript.

References

1. Bauschke, H.H., Bui, M.N., Wang, X.: On sums and convex combinations of projectors onto convex sets. J. Approx. Theory **242**, 31–57 (2019)
2. Beck, A.: First-Order Methods in Optimization, vol. 25. SIAM, Philadelphia (2017)
3. Bergen, G.: A fast and robust GJK implementation for collision detection of convex objects, Tech. report, Department of Mathematics and Computing Science, Eindhoven University of Technology (1999)
4. Borwein, J.M., Lewis, A.S.: Convex Analysis and Nonlinear Optimization: Theory and Examples. CMS Books in Mathematics, Canadian Mathematical Society (2000)
5. Cameron, S.: Enhancing GJK: computing minimum and penetration distances between convex polyhedra, vol. 3112–3117 (1997)
6. Chang, L., Qiao, H., Wan, A., Keane, J.: An improved Gilbert algorithm with rapid convergence. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3861–3866 (2006)
7. Chen, Y., Ye,X.: Projection onto a Simplex, CoRR, abs/1208.4873
8. Condat, L.: Fast projection onto the simplex and the ℓ_1 ball. Math. Program. **158**, 575–585 (2016)
9. Dai, Y.H.: Fast algorithms for projection on an ellipsoid. SIAM J. Optim. **16**, 986–1006 (2006)
10. Dax, A.: A new class of minimum norm duality theorems. SIAM J. Optim. **19**, 1947–1969 (2009)
11. Defazio, A., Bach, F., Lacoste-Julien, S.: Saga: a fast incremental gradient method with support for non-strongly convex composite objectives. In: Advances in Neural Information Processing Systems (2014)
12. Duchi, J., Shalev-Shwartz, S., Singer, Y., Chandra, T.: Efficient projections onto the ℓ_1-ball for learning in high dimensions. In: Proceedings of the 25th ACM International Conference on Machine Learning, pp. 272–279 (2008)
13. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval Res. Logist. Q. **3**, 95–110 (1956)
14. Gabidullina, Z.R.: The problem of projecting the origin of euclidean space onto the convex polyhedron, arXiv:1605.05351
15. Gaines, B.R., Kim, J., Zhou, H.: Algorithms for fitting the constrained lasso. J. Comput. Graph. Stat. **27**(4), 861–871 (2018)
16. Garber, D., Hazan, E.: Faster rates for the frank-wolfe method over strongly-convex sets. In: ICML, 541–549 (2015)
17. Gilbert, E.G.: An iterative procedure for computing the minimum of a quadratic form on a convex set. SIAM J. Control **4**, 61–80 (1966)
18. Gilbert, E.G., Johnson, D.W., Keerthi, S.S.: A fast procedure for computing the distance between complex objects in three-dimensional space. IEEE Trans. Robot. Autom. **4**, 193–203 (1988)
19. Gilbert, E.G., Foo, C.-P.: Computing the distance between general convex objects in three-dimensional space. IEEE Trans. Robot. Autom. **6**, 53–61 (1990)

20. Hiriart-Urruty, J.B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms, I and II, Grundlehren Math. Wiss. 305 and 306. Springer, Berlin (1993)
21. Jaggi, M.: Revisiting Frank–Wolfe: projection-free sparse convex optimization. In: ICML, vol. 1, pp. 427–435 (2013)
22. Lacoste-Julien, S., Jaggi, M.: On the global linear convergence of Frank–Wolfe optimization variants. In: Advances in Neural Information Processing Systems, pp. 496–504 (2015)
23. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: A fast iterative nearest point algorithm for support vector machine classifier design. IEEE Trans. Neural Netw. **11**, 124–136 (2000)
24. Kurzhanskiy, A.A., Varaiya, P.: Ellipsoidal Toolbox, Tech. Report EECS-2006-46, EECS, UC Berkeley (2006)
25. Luenberger, D.G.: Optimization by Vector Spaces Method. Wiley, New York (1969)
26. Martin, S.: Training support vector machines using Gilbert's algorithm. In: The 5th IEEE International Conference on Data Mining (ICDM), pp. 306–313 (2005)
27. Mitchell, B.F., Demyanov, V.F., Malozemov, V.N.: Finding the point of a polyhedron closest to the origin. SIAM J. Control Optim. **12**, 19–26 (1974)
28. Mordukhovich, B.S., Nam, N.M.: Limiting subgradients of minimal time functions in Banach spaces. J. Glob. Optim. **46**, 615–633 (2010)
29. Nam, N.M., An, N.T., Rector, R.B., Sun, J.: Nonsmooth algorithms and Nesterov smoothing technique for generalized Fermat–Torricelli problems. SIAM J. Optim. **24**(4), 1815–1839 (2014)
30. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**, 127–152 (2005)
31. Nesterov, Y.: Introductory lectures on convex optimization: a basic course, Appl. Optim. 87, Kluwer, Boston (2004)
32. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $O\left(\frac{1}{k^2}\right)$. Dokl. Akad. Nauk SSSR **269**, 543–547 (1983)
33. Nirenberg, L.: Functional Analysis. Academic Press, New York (1961)
34. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
35. Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Technical report, INRIA, hal-0086005 (2013)
36. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc.: Ser. B (Methodological)) **58**(1), 267–288 (1996)
37. Tibshirani, R., Taylor, J.: The solution path of the generalized lasso. Ann. Stat. **39**(3), 1335–1371 (2011)
38. Tuy, H.: Convex Analysis and Global Optimization: Nonconvex Optimization and Its Applications. Kluwer, Dordrecht (1998)
39. Wolfe, P.: Finding the nearest point in a polytope. Math. Programm. **11**, 128–149 (1976)
40. Won, J.H., Xu, J., Lange, K.: Projection onto Minkowski sums with application to constrained learning. In: International Conference on Machine Learning, pp. 3642–3651 (2019)
41. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc.: Ser. B (Statistical Methodology) **68**(1), 49–67 (2006)
42. Yuan, L., Liu, J., Ye, J.: Efficient methods for overlapping group lasso. In: Advances in Neural Information Processing Systems, pp. 352–360 (2011)