kFW: A Frank-Wolfe style algorithm with stronger subproblem oracles

Lijun Ding LD446@CORNELL.EDU
Cornell University

Jicong Fan FANJICONG@CUHK.EDU.CN
CUHZ (Shenzhen)

Madeleine Udell UDELL@CORNELL.EDU
Cornell University

Abstract

This paper proposes a new variant of Frank-Wolfe (FW), called kFW. Standard FW suffers from slow convergence: iterates often zig-zag as update directions oscillate around extreme points of the constraint set. The new variant, kFW, overcomes this problem by using two stronger subproblem oracles in each iteration. The first is a k linear optimization oracle (kLOO) that computes the k best update directions (rather than just one). The second is a k direction search (kDS) that minimizes the objective over a constraint set represented by the k best update directions and the previous iterate. When the problem solution admits a sparse representation, both oracles are easy to compute, and kFW converges quickly for smooth convex objectives and several interesting constraint sets: kFW achieves finite $\frac{4L_f^3 D^4}{\gamma\delta^2}$ convergence on polytopes and group norm balls, and linear convergence on spectrahedra and nuclear norm balls. Numerical experiments validate the effectiveness of kFW and demonstrate an order-of-magnitude speedup over existing approaches.

1. Introduction

We consider the following optimization problem with decision variable x:

$$\begin{array}{ll} \text{minimize} & f(x) := g(\mathcal{A}x) + \langle c, x\rangle \\ \text{subject to} & x \in \Omega. \end{array} \tag{1}$$

The constraint set $\Omega \subset \mathbf{E}$ is a convex and compact subset of a finite dimensional Euclidean space $\mathbf{E}$ and has diameter D[1]. The map $\mathcal{A} : \mathbf{E} \to \mathbf{F}$ is linear, where $\mathbf{F}$ is another finite dimensional Euclidean space. We equip both spaces $\mathbf{E}$ and $\mathbf{F}$ with real inner products denoted as $\langle\cdot,\cdot\rangle$. The vector c is in $\mathbf{E}$. The function $g : \mathbf{F} \to \mathbb{R}$ is convex and L_g-smooth[2]. The smoothness of g implies f is L_f-smooth for some $L_f > 0$. For ease of exposition, we assume Problem (1) admits a unique solution.[3]

Applications. The optimization problem (1) appears in a wide variety of applications, such as sparse vector recovery [5], group-sparse vector recovery [36], combinatorial problems [27], submodular optimization [2, 37], and low rank matrix recovery problems [26, 34].

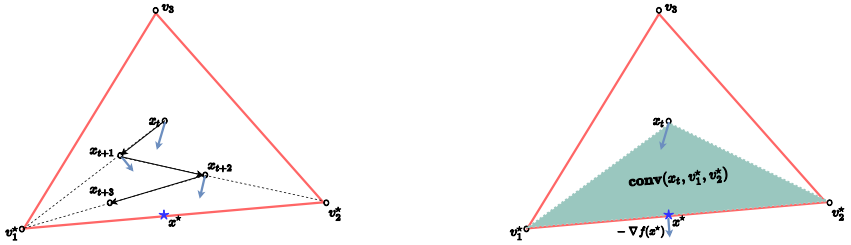
1. The diameter of Ω is defined as $\sup_{x,y\in\Omega}\|x-y\|$, where $\|\cdot\|$ is the norm inducded by the inner product.
2. That is, the gradient of g is L_g-Lipschitz continuous with respect to the norm $\|\cdot\|$.
3. The main result Theorem 2 remains valid for multiple optimal solution setting after minor adjustments, see Section D. Note that from [10, Corollary 3.5], the solution is indeed unique for almost all c.

Frank-Wolfe and two subproblems. In many modern high-dimensional applications, Euclidean projection onto the set Ω is challenging. Hence the well-known projected gradient (PG) method and its acceleration version (APG) are not well suited for (1). Instead, researchers have turned to projection-free methods, such as the Frank-Wolfe algorithm (FW) [15] shown in Algorithm 1, also known as the conditional gradient method [31, Section 6]. The linear optimization oracle can be computed efficiently for many interesting constraint sets Ω even when projection is prohibitively expensive. These sets include the probability simplex, the ℓ_1 norm ball, and many more polytopes arising from combinatorial optimization, the spectrahedron $\mathcal{SP}^n = \{X \in \mathbb{S}^n_+ \mid \mathbf{tr}(X) = 1\}$, and the unit nuclear norm ball $\mathbf{B}_{\|\cdot\|_{\text{nuc}}} = \{X \mid \|X\|_{\text{nuc}} \leq 1\}$. We refer the reader to [25, 29] for further examples. Line search is easy to implement using a closed formula for quadratic f, or bisection.

Algorithm 1 Frank-Wolfe with line search

Input: initialization $x_0 \in \mathbf{E}$
for $t = 1, 2, \ldots,$ **do**
  Linear optimization oracle (LOO): Compute $v_t \in \arg\min_{v \in \mathbf{E}} \langle v, \nabla f(x) \rangle$.
  Line search: solve $\hat{\eta} = \arg\min_{\eta \in [0,1]} f(\eta x_t + (1-\eta) v_t)$ and set $x_{t+1} = \hat{\eta} x_t + (1-\hat{\eta}) v_t$.
end for

Slow convergence of FW and the Zigzag. However, FW is known to be slow in both theory and practice, reaching an accuracy of $\mathcal{O}(\frac{1}{t})$ after t iterations. This slow convergence is often described pictorially by the *Zigzag phenomenon* depicted in Figure 1(*a*)subfigure. The Zigzag phenomenon occurs when the optimal solution $x_\star$ of (1) lies on the boundary of Ω and is a convex combination of $r_\star$ many extreme points $v_1^\star, \ldots, v_{r_\star}^\star \in \Omega$: $x_\star = \sum_{i=1}^{r_\star} \lambda_i^\star v_i^\star, \quad \lambda_i^\star > 0,$ and $\sum_{i=1}^{k} \lambda_i^\star = 1$ (In Figure 1(*a*)subfigure, $r_\star = 2$). When Ω is a polytope, the LOO will alternate between the extreme points $v_i^\star$s and the line search updates the estimate of $\lambda_i^\star$ slowly as the iterate approaches to $x_\star$. A similar Zigzag occurs for other sets such as the spectrahedron and nuclear norm ball. A long line of work has explored methods to reduce the complexity of FW using LOO and line search alone [16, 18, 19, 22, 28, 29].



(*a*) Zig-Zag phenomenon of FW (*b*) Optimization over $\mathbf{conv}(x_t, v_1^\star, v_2^\star)$

Figure 1: The Zigzag phenomenon and optimization over $\mathbf{conv}(x_t, v_1^\star, v_2^\star)$. Here, the solution $x_\star$ is a convex combination of $v_1^\star$ and $v_2^\star$, and $r_\star = 2$. The grey arrows are the negative gradients $-\nabla f$.

Our key insight: overcoming zigzags with kFW. Our first observation is that the sparsity $r_\star$ is expected to be small for most large scale applications mentioned. For example, the sparsity is the number of nonzeros in sparse vector recovery, the number of nonzero groups in group-sparse vector recovery, and the rank in low rank matrix recovery. Next, note that from the optimality condition (also see Figure 1(*b*)subfigure), the gradient $\nabla f(x_\star)$ in this case has the smallest inner product

with $v_1^\star, \ldots, v_{r_\star}^\star$ among all $v \in \Omega$. Also, for small $r_\star$, we can solve $\min_{x \in \mathbf{conv}(x_t, v_1^\star, \ldots, v_{r_\star}^\star)} f(x)$ efficiently [4] to obtain the solution $x_\star$. Hence, our key insight to overcome the Zigzag is simply

Compute all extreme points $v_i^\star$ *that minimize* $\langle \nabla f(x_\star), v \rangle$
and solve the smaller problem $\min_{x \in \mathbf{conv}(x_t, v_1^\star, \ldots, v_{r_\star}^\star)} f(x)$.

This insight leads us to define a new algorithmic ways to choose extreme points and define a smaller convex search set, which we call kLOO and kDS. For polytope Ω, they are defined as

- k linear optimization oracle (kLOO): for any $y \in \mathbb{R}^n$, compute the k extreme points $v_1, \ldots, v_k$ (k best directions) with the smallest k inner products $\langle v, y \rangle$ among all extreme points v of Ω.
- k direction search (kDS): given $w, v_1, \ldots, v_k \in \Omega$, output $x_{k\text{DS}} = \arg\min_{x \in \mathbf{conv}(w, v_1, \ldots, v_k)} f(x)$.

In connection with FW, kLOO and kDS can be considered stronger subproblem oracles compared to LOO and line search respectively. Combining the two subproblem oracles, we arrive at a new variant of the Frank-Wolfe algorithm: kFW, presented in Algorithm 2. In Section 2, we show that the two subproblems can actually be efficiently solved over many polytopes (for small k). Moreover, we redefine kLOO and kDS to incorporate the situation where k best extreme points are not well-defined for sets such as group norm ball, spectrahedron, and nuclear norm ball, yet sparsity structure still persists. Finally, we note that with our terminology, 1FW is the same as FW. Hence our main results, Theorem 2, give new insight into the fast convergence of FW when $r_\star = 1$.

Algorithm 2 kFW

Input: initialization $x_0 \in \Omega$, and an integer $k > 0$
for $t = 1, 2, \ldots,$ **do**
  Compute kLOO with input $\nabla f(x_t)$
  Compute kDS with input consisting of x_t and the output of kLOO, and output $x_{t+1} = x_{k\text{DS}}$
end for

Computational efficiency of kFW. Here we summarize the computational efficiency of kFW in terms of its per iteration cost, iteration complexity, and storage:

- *Per iteration cost:* For many important cases displayed in Section 2, kFW admits efficient subproblem oracles.
- *Iteration complexity:* kFW achieves the same $\mathcal{O}(1/t)$ convergence rate of FW. Under additional regularity conditions, it achieves nonexponential finite convergence over the polytope and group norm ball and linear convergence over the spectrahedron and nuclear norm ball, as shown in Theorem 2. These results are beyond the reach of many FW variants [16, 19, 22, 29].
- *Storage:* The storage required by kFW is $\mathcal{O}(kn)$, needed to store the k best directions computed in each step, while the pairwise step, away step, and fully corrective step based FW [29] require $\mathcal{O}(\min(tn, n^2))$ storage to accumulate vectors of dimension n.

A comparison of kFW with FW, away-step FW [22], and Fully corrective FW (FCFW) [25, Algorithm 4] is shown in Table 2 in Section A.1 in the appendix.

Paper Organization. In Section 2, we explain how to solve the two subproblems over a polytope Ω, and how to extend the idea to group norm ball, spectrahedron, and nuclear norm ball. In Section 3, we describe a few analytical conditions, and then present the faster convergence guarantees of kFW under these conditions for the polytope, unit group norm ball, spechedron, and unit nuclear norm ball. We demonstrate the effectiveness of kFW numerically in Section 4.

4. Here $\mathbf{conv}(v_1^\star, \ldots, v_{r_\star}^\star)$ is the convex hull of $v_1, \ldots, v_{r_\star}$.

2. Stronger subproblem oracles for polytopes and beyond

2.1. Stronger subproblem oracles for polytopes

Solving kLOO. Computing a LOO can be NP-hard for some constraint sets Ω: for example, the 0 - 1 knapsack problem can be formulated as linear optimization over an appropriate polytope. Hence we should not expect that we can compute a kLOO efficiently without further assumptions on the polytope $\Omega \subset \mathbb{R}^n$. Since many polytopes come from problems in combintorics, for these polytopes, computing a kLOO is equivalent to computing the k best solutions to a problem in the combinatorics literature, and polynomial time algorithms are available for many polytopes [14, 23, 30, 33]. We present the time complexity of computing kLOO for many interesting problems in Table 3 in the appendix. Moreover, we note that examples of efficient kLOO, time complexity no more than k times of LOO, are abundant: the probability simplex, the ℓ_1 norm ball, the spanning tree polytope [12], the Birkhoff polytope, [33], and the flow polytope of a directed acyclic graph [13] all admit efficient kLOO. More details of each example and its application are in Section A.4 in the appendix.

k direction search. The k direction search problem optimizes the objective $f(x)$ over $x \in \mathbf{conv}(w, v_1, \ldots, v_k) = \{\sum_{i=1}^k \lambda_i v_i + \eta w \mid (\eta, \lambda) \in \Delta^{k+1}\}$. We parametrize this set by $(\eta, \lambda) \in \Delta^{k+1}$ and employ the accelerated projected gradient method (APG) to solve

$$\min_{(\eta,\lambda)\in\Delta^{k+1}} f\left(\sum_{i=1}^k \lambda_i v_i + \eta w\right). \tag{2}$$

The constraint set here is a $k+1$ dimensional probability simplex; projection onto this set requires time $\mathcal{O}(k \log k)$ [6]. Hence for small k, we can solve (2) efficiently. We recover the output $x_{k-\text{DS}} = \sum_{i=1}^k \lambda_i^\star v_i + \eta^\star w$ of kDS from the optimal solution $(\eta^\star, \lambda^\star)$ of (2).

2.2. Stronger subproblem oracle for nonpolytope Ω

Due to space limit, we explain kLOO and kDS for (i) group norm ball, (ii) nuclear norm ball, and (iii) spectrahedron in Section A.2 and A.3 respectively in the appendix. A summary of kLOO and kDS for these sets appears in Table 4 and 5 in the appendix respectively. The kFW algorithm for unit group norm ball, spectrahedron, and unit nuclear norm ball is presented as Algorithm 2 as well.

3. Theoretical guarantees

In this section, we first present a few definitions and conditions required to state our results: sparsity measure, strict complementarity, and quadratic growth. Due to space limit, we only presents conditions concerning polytope. Strict complementarity is necessary for robustness of the sparsity (Example B.1) and actually implies quadratic growth (Theorem 4). More details are in the appendix.

Definition 1 *The sparsity measure, strict complementarity, and quadratic growth are defined as:*

- *Sparsity measure $r_\star$: the number of extreme points of the smallest face $\mathcal{F}(x_\star)$ containing $x_\star$.*
- *Strict complementarity (SC) and its measure δ: Problem (1) admits SC if it has a unique solution $x_\star \in \partial\Omega$ and $-\nabla f(x_\star) \in$ relint$(N_\Omega(x_\star))$[5] The complementarity measure δ is the gap between the inner products of $x_\star$ and the elements of the* complementary set *$\mathcal{F}^c(x_\star)$, the convex hull of all vertices not in $\mathcal{F}(x_\star)$: $\delta = \min\{\langle u, \nabla f(x_\star)\rangle - \langle x_\star, \nabla f(x_\star)\rangle \mid u \in \mathcal{F}^c(x_\star) \subset \Omega\}$.*
- *Quadratic growth (QG): Problem (1) admits QG with parameter $\gamma > 0$ if it has a unique solution $x_\star$ and for all $x \in \Omega$, $f(x) - f(x_\star) \geq \gamma\|x - x_\star\|^2$.*

5. Here $\partial\Omega$ is the topological boundary of Ω under standard topology of $\mathbf{E}$. The set $N_\Omega(x_\star)$ is the normal cone of Ω at $x_\star$, i.e. $N_\Omega(x_\star) = \{y \mid \langle y, x\rangle \leq \langle y, x_\star\rangle, \forall x \in \Omega\}$, and relint$(\cdot)$ is the relative interior.

Our following theorem states that kFW never requires more iterations than FW and terminates in finite time for polytope and group norm ball, while converges linearly for nuclear norm ball and spectrahedron because any neighborhood of $X_\star$ contains infinitely many matrices with rank $r_\star$. Proofs are deferred to Section B.4.

Theorem 2 *Suppose f is L_f-smooth and convex and Ω is convex compact with diameter D. Then for any $k \geq 1$ and for all $t \geq 1$, the iterate x_t in kFW (Algorithm 2) satisfies $f(x_t) - f(x_\star) \leq \frac{2L_f D^2}{t}$. Moreover, suppose Problem (1) satisfies strict complementarity and quadratic growth, and $k \geq r_\star$. If the constraint set Ω is a polytope or a unit group norm ball, then the gap $\delta > 0$ and kFW finds $x_\star$ in at most $T + 1$ iterations, where T is*

$$T = \frac{4L_f^3 D^4}{\gamma \delta^2}. \tag{3}$$

If the constraint set is the spechedron or the unit nuclear norm ball, the gap $\delta > 0$ and kFW satisfies that for any $t \geq T_1 := \frac{72L_f^3}{\gamma\delta^2}$, $f(X_{t+1}) - f(X_\star) \leq \left(1 - \min\left\{\frac{\gamma}{4L_f}, \frac{\delta}{12L_f}\right\}\right)(f(X_t) - f(X_\star))$.

4. Numerics

We compare our method kFW with FW, away-step FW (awayFW) [22], pairwise FW (pairFW)[29], DICG [19], and blockFW [1] for the Lasso, support vector machine (SVM), group Lasso, and matrix completion problems on synthetic data. Details about experimental settings appear in the Appendix E. All algorithms terminate when the relative change of the objective is less than 10^{-6} or after 1000 iterations. As shown in Figure 2, kFW converges in many fewer iterations than other methods. Table 1 shows that kFW also converges faster in wall-clock time, with one exception (blockFW in matrix completion). Note that blockFW is sensitive to the step size while kFW has no step size to tune.

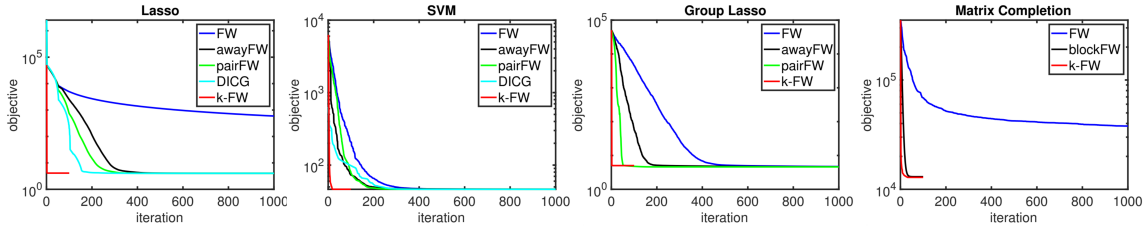


Figure 2: kFW vs. FW and its variants (more numerics can be found in Appendix E).

Table 1: Computation time (seconds): the algorithms terminate when the relative change of the objective $< 10^{-6}$ or after 1000 iterations. Sign " -" means the algorithm is not suited to the problem.

	FW	awayFW	pairFW	DICG	blockFW	kFW
Lasso	>14	7	6	10	-	**0.5**
SVM	6	4.5	2.9	2.5	-	**0.6**
Group Lasso	17	6	1.8	-	-	**0.3**
Matrix completion	>180	-	-	-	1.8	4.8

5. Conclusion and discussion

This paper presented a new variant of FW, kFW, that takes advantage of sparse structure in problem solutions to offer much faster convergence than other variants of FW, both in theory and in practice. kFW avoids the Zigzag phenomenon by optimizing over a convex combination of the previous iterate and k extreme points of the constraint set, rather than one, at each iteration. The method relies on the ability to efficiently compute these k extreme points (kLOO) and to compute the update (kDS), which we demonstrate for a variety of interesting problems. We expect the core ideas that undergird kFW can be generalized to a wide variety of atomic sets in addition to those considered in this paper.

References

[1] Zeyuan Allen-Zhu, Elad Hazan, Wei Hu, and Yuanzhi Li. Linear convergence of a frank-wolfe type algorithm over trace-norm balls. In *Advances in Neural Information Processing Systems*, pages 6191–6200, 2017.

[2] Francis Bach et al. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in Machine Learning*, 6(2-3):145–373, 2013.

[3] Heinz H Bauschke, Jonathan M Borwein, and Wu Li. Strong conical hull intersection property, bounded linear regularity, jameson's property (g), and error bounds in convex optimization. *Mathematical Programming*, 86(1):135–160, 1999.

[4] Amir Beck and Shimrit Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1-2):1–27, 2017.

[5] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[6] Yunmei Chen and Xiaojing Ye. Projection onto a simplex. *arXiv preprint arXiv:1101.6081*, 2011.

[7] Kenneth L Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):1–30, 2010.

[8] Lijun Ding and Madeleine Udell. On the regularity and conditioning of low rank semidefinite programs. *arXiv preprint arXiv:2002.10673*, 2020.

[9] Lijun Ding, Yingjie Fei, Qiantong Xu, and Chengrun Yang. Spectral frank-wolfe algorithm: Strict complementarity and linear convergence. *arXiv print arXiv:2006.01719*, 2020.

[10] Dmitriy Drusvyatskiy and Adrian S Lewis. Generic nondegeneracy in convex optimization. *Proceedings of the American Mathematical Society*, pages 2519–2527, 2011.

[11] Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

[12] David Eppstein. Finding the k smallest spanning trees. In *Scandinavian Workshop on Algorithm Theory*, pages 38–47. Springer, 1990.

[13] David Eppstein. Finding the k shortest paths. *SIAM Journal on computing*, 28(2):652–673, 1998.

[14] David Eppstein. k-best enumeration. *arXiv preprint arXiv:1412.5075*, 2014.

[15] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[16] Robert M Freund, Paul Grigas, and Rahul Mazumder. An extended frank–wolfe method with "in-face" directions, and its application to low-rank matrix completion. *SIAM Journal on optimization*, 27(1):319–346, 2017.

[17] Dan Garber. On the convergence of projected-gradient methods with low-rank projections for smooth convex minimization over trace-norm balls and related problems. *arXiv preprint arXiv:1902.01644*, 2019.

[18] Dan Garber and Elad Hazan. Faster rates for the frank-wolfe method over strongly-convex sets. In *32nd International Conference on Machine Learning, ICML 2015*, 2015.

[19] Dan Garber and Ofer Meshi. Linear-memory and decomposition-invariant linearly convergent conditional gradient algorithm for structured polytopes. In *Advances in neural information processing systems*, pages 1001–1009, 2016.

[20] Tom Goldstein, Christoph Studer, and Richard Baraniuk. A field guide to forward-backward splitting with a FASTA implementation. *arXiv eprint*, abs/1411.3406, 2014. URL http://arxiv.org/abs/1411.3406.

[21] Tom Goldstein, Christoph Studer, and Richard Baraniuk. FASTA: A generalized implementation of forward-backward splitting, January 2015. http://arxiv.org/abs/1501.04979.

[22] Jacques Guélat and Patrice Marcotte. Some comments on wolfe's 'away step'. *Mathematical Programming*, 35(1):110–119, 1986.

[23] Horst W Hamacher and Maurice Queyranne. K best solutions to combinatorial optimization problems. *Annals of Operations Research*, 4(1):123–143, 1985.

[24] Christoph Helmberg and Franz Rendl. A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10(3):673–696, 2000.

[25] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th international conference on machine learning*, number CONF, pages 427–435, 2013.

[26] Martin Jaggi and Marek Sulovskỳ. A simple algorithm for nuclear norm regularized problems. 2010.

[27] Armand Joulin, Kevin Tang, and Li Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *European Conference on Computer Vision*, pages 253–268. Springer, 2014.

[28] Simon Lacoste-Julien and Martin Jaggi. An affine invariant linear convergence analysis for frank-wolfe algorithms. *arXiv preprint arXiv:1312.7864*, 2013.

[29] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Advances in Neural Information Processing Systems*, pages 496–504, 2015.

[30] Eugene L Lawler. A procedure for computing the k best solutions to discrete optimization problems and its application to the shortest path problem. *Management science*, 18(7):401–405, 1972.

[31] Evgeny S Levitin and Boris T Polyak. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 6(5):1–50, 1966.

[32] Conrado Martínez and Salvador Roura. Optimal sampling strategies in quicksort and quickselect. *SIAM Journal on Computing*, 31(3):683–705, 2001.

[33] Katta G Murthy. An algorithm for ranking all the assignments in order of increasing costs. *Operations research*, 16(3):682–687, 1968.

[34] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[35] Suvrit Sra. Fast projections onto $\ell_{1,q}$-norm balls for grouped feature selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 305–317. Springer, 2011.

[36] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[37] Song Zhou, Swati Gupta, and Madeleine Udell. Limited memory Kelley's method converges for composite convex and submodular objectives. In *Advances in Neural Information Processing Systems*, pages 4414–4424, 2018.

[38] Zirui Zhou and Anthony Man-Cho So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728, 2017.

Appendix A. Table and Procedures for Section 2

A.1. Comparison of iteration complexity with FW variants for polytope Ω

Here we present a table for comparing kFW with FW, away-step FW [22], and Fully corrective FW (FCFW) [25, Algorithm 4].

Table 2: Comparison of kFW, FW, FW with away step, and FCFW for Problem (1), smooth convex optimization over a constraint set Ω in a n-dimensional Euclidean space. We display the per iteration computation (per iter. comp.), storage, faster rate (compared to $\mathcal{O}(\frac{1}{t})$ rate) under the condition on Ω, extra conditions on Problem (1) to achieve the faster rate (Ex. Cond.), and the reference providing the proof of the rate. Even without the extra conditions listed in the table, all algorithms admit a $\mathcal{O}(\frac{1}{t})$ convergence rate (see [25] and Theorem 2). Here $t \wedge n = \min(t, n)$. Definitions of the sparsity measure $r_\star$, strict complementarity (str. comp.), and quadratic growth (q.g.) can be found in Section 3.

Algorithm	Per iter. comp.	Storage	Rate and Ω Shape	Ex. Cond.	Reference
FW	LOO, 1DS	$\mathcal{O}(n)$	finite polytope, group norm ball	str. comp., q.g., and $r_\star = 1$	Theorem 2
			linear spectrahedron $\mathbf{B}_{\|\cdot\|_{\text{nuc}}}$		Theorem 2
Away-step FW	LOO, 1DS, and $t \wedge n$ inner products	$\mathcal{O}(n(t \wedge n))$	linear polytope	q.g.	[29]
FCFW	LOO, and $(t \wedge n)$DS	$\mathcal{O}(n(t \wedge n))$	linear polytope	q.g	[29]
kFW	kLOO, and kDS	$\mathcal{O}(kn)$	finite polytope, group norm ball	str. comp., q.g., and $k \geq r_\star$	Theorem 2
			linear spectrahedron $\mathbf{B}_{\|\cdot\|_{\text{nuc}}}$		Theorem 2

A.2. kLOO and kDS for Nuclear norm ball

We now define kLOO and kDS for the unit nuclear norm ball $\mathbf{B}_{\|\cdot\|_{\text{nuc}}} = \{X \in \mathbb{R}^{n_1 \times n_2} \mid \|X\|_{\text{nuc}} \leq 1\}$, where $\|X\|_{\text{nuc}} = \sum_{i=1}^{\min(n_1,n_2)} \sigma_i(X)$, the sum of singular values.

kLOO. Given an input matrix $Y \in \mathbb{R}^{n_1 \times n_2}$, define the k best directions of the linearized objective $\min_{V \in \Omega(\alpha)} \langle V, Y \rangle$ to be the pairs $(u_1, v_1), \ldots, (u_k, v_k)$, the top k left and right orthonormal singular vectors of Y. Collect the output as $U = [u_1, \ldots, u_k] \in \mathbb{R}^{n_1 \times k}$ and $V = [v_1, \ldots, v_k] \in \mathbb{R}^{n_2 \times k}$.

kDS. Take as inputs $W \in \mathbf{B}_{\|\cdot\|_{\text{nuc}}}$ and $(U, V) \in \mathbb{R}^{n_1 \times k} \times \mathbb{R}^{n_2 \times k}$ with orthonormal columns. Inspired by [24], we consider the spectral convex combinations of W and $u_i v_i^\top$ instead of just convex combination:

$$X = \eta W + U S V^\top \quad \text{where} \quad \eta \geq 0,\ \eta + \|S\|_{\text{nuc}} \leq 1.$$

Next, we minimize the objective $f(X)$ parametrized by $(\eta, S) \in \mathbb{R}^{1+k^2}$ to obtain $X_{k\text{DS}}$:

$$\text{minimize} \quad f(\eta W + USV^\top) \quad \text{subject to} \quad \eta + \|S\|_{\text{nuc}} \le 1, \eta \ge 0.$$

Again, we use APG to solve this problem. Projection requires singular value decomposition of a k^2 matrix, which is tolerable for small k. (See Section A.8 for details.)

A.3. kLOO and kDS for Spectrahedron

Recall a matrix $A \in \mathbb{S}^n$ is positive semidefinite if all its eigenvalues are nonnegative. We denote a positive semidefinite matrix as $A \succeq 0$ or $A \in \mathbb{S}^n_+$. We write the eigenvalues of a symmetric matrix $A \in \mathbb{S}^n$ as $\lambda_1(A) \ge \cdots \ge \lambda_n(A)$.

We now define kLOO and kDS for the spectrahedron $\mathcal{SP}^n = \{X \in \mathbb{S}^n \mid X \succeq 0, \, \mathbf{tr}(X) = 1\}$.

kLOO. Given a input matrix $Y \in \mathbb{S}^n$, define the k best directions of the linearized objective $\min_{V \in \mathcal{SP}^n} \langle V, Y \rangle$ as the bottom k eigenvectors of Y, the eigenvectors corresponding to the k smallest eigenvalues. Call these vectors $v_1, \ldots, v_n$ and collect the output as $V = [v_1, \ldots, v_k] \in \mathbb{R}^{n \times k}$.

kDS. Take as inputs $W \in \mathcal{SP}^n$ and $V = [v_1, \ldots, v_k] \in \mathbb{R}^{n \times r}$ with orthonormal columns. Instead of convex combinations of W and $v_i v_i^\top$, we consider a spectral variant inspired by [24]:

$$X = \eta W + VSV^\top \quad \text{where} \quad \eta \ge 0, \; S \in \mathbb{S}^k_+, \; \eta + \mathbf{tr}(S) = 1.$$

We minimize the objective $f(X)$ over this constraint set to obtain the solution $X_{k\text{DS}}$ to kDS:

$$\text{minimize} \quad f(\eta W + VSV^\top) \quad \text{subject to} \quad \eta \ge 0, \; S \in \mathbb{S}^k_+, \; \text{and } \eta + \mathbf{tr}(S) = 1.$$

Again, we use APG to solve this problem. Projection onto the constraint set requires eigenvalue decomposition (EVD) of a k^2 matrix, which is tolerable for small k. (See Section A.8 in the appendix.)

A.4. kLOO of combinatorical optimization

In this section, we present Table 3 of the computational complexity of finding the k best solution for combinatorical optimizations. In our setting, the k best solution corresponds to the k best directions of kLOO. We then point out those kLOO that can be efficiently computed.

Efficient kLOO. Though kLOO can be computed in polynomial times, it is unfortunately that for some polytopes, the time required to compute a kLOO grows superlinearly in k even if assuming $k \le n$. Hence we restrict our attention to special structured polytopes for which the time complexity of kLOO is no more than k times the complexity of LOO.

Our primary example is the probability simplex $\Delta^n = \mathbf{conv}(\{e_i\}_{i=1}^n)$ in $\mathbb{R}^n$. Since the vertices of Δ^n are the coordinate vectors $e_i, i = 1, \ldots, n$, the inner product of vertex e_i with a vector $y \in \mathbb{R}^n$ is $\langle y, e_i \rangle = y_i$. Hence in this case, kLOO with input $y \in \mathbb{R}^n$ simply outputs the coordinate vectors corresponding to the smallest k values of y. Using a binary heap of k nodes, we can scan through the entries of y and update the heap to keep the k smallest entries seen so far and their indices. Since each heap update takes time $\mathcal{O}(\log k)$, the time to compute kLOO is $\mathcal{O}(n \log k)$. A more sophisticated procedure called *quickselect* improves the time to $\mathcal{O}(n + k)$ [32], [14, Section 2.1]. Let us now list other examples admitting efficient kLOO:

- The ℓ_1 norm ball $\{x \in \mathbb{R}^n \mid \sum_{i=1}^n |x_i| \leq \alpha\}$ admits a kLOO with time complexity $\mathcal{O}(n+k)$ by simply considering finding the k largest elements among $2n$ elements.
- The spanning tree polytope of an undirected graph $G(V, E)$ in $\mathbb{R}^{|E|}$ admits a kLOO with time complexity $\mathcal{O}(m \log n + k^2)$, where $m = |E|$ and $n = |V|$ [12].
- The Birkhoff polytope, the convex hull of permutation matrices in $\mathbb{R}^{n \times n}$, admits a kLOO with time complexity $\mathcal{O}(kn^3)$ [33]
- The flow polytope of a directed acyclic graph $G(V, E)$ in $\mathbb{R}^{|E|}$ admits a kLOO with time complexity $\mathcal{O}(m + n \log n + k \min(n, k)^{1/2})$, where $m = |E|$ and $n = |V|$ [13].

Optimization over the probability simplex is useful for fitting support vector machines [7, Problem (24)]. The ℓ_1 norm ball plays a key role in sparse signal recovery [5]. The flow polytope appears in applications in video-image co-localization [27].

Table 3: The time complexity of kLOO for different combinatorical problems. The matroid $M = (E, \mathcal{I})$ consists of the ground set E with n elements and the set of bases $\mathcal{I}$. The polytope is the convex hull of all bases in $[0, 1]^n$. The quantity α is the complexity of checking independence of a set. Here $r(M)$ is the rank of the matroid M. The $s-t$ cut is for a directed graph with n nodes, m edges, a source node s, and a sink node t. For each $s-t$ cut, a partition S, S^c of the vertex set with $s \in S$ and $t \in S^c$, we define its cut point as a vector in $[-1, 1]^m$ that has entry 1 for an edge from S to S^c, and an entry -1 for an edge from S^c to S. The $s-t$ cut polytope is the convex hull of all cut points in $[-1, 1]^m$. The path polytope consider all simple path from s to t for a directed acyclic graph with n nodes and m edges. The polytope is then the convex hull of all simple path point in $[0, 1]^m$. For an undirected graph with n nodes and m edges, the spanning tree polytope is the convex hull of all spanning tree in $[0, 1]^m$.

Polytope name	LOO complexity	kLOO complexity
Probability simplex	$\mathcal{O}(n)$	$\mathcal{O}(n+k)$ [32]
Polytope of bases of a matroid M	$\mathcal{O}(n \log n, n\alpha)$	$\mathcal{O}(n \log n + knr(M)\alpha)$ [23]
The Birkhoff polytope	$\mathcal{O}(n^3)$	$\mathcal{O}(kn^3)$ [33]
$s-t$ Cut Polytope (Directed Graph)	$\mathcal{O}(nm \log n)$	$K\mathcal{O}(n^4)$ [23]
$s-t$ path Polytope(DAG)	$\mathcal{O}(m + n \log n)$	$\mathcal{O}(m + n \log n + kn)$ [13]
Spanning tree Polytope	$\mathcal{O}(m + n \log n)$	$\mathcal{O}(m \log n + k \min(n, k)^{1/2})$ [12]

A.5. Tables for kLOO and kDS

This section presents Table 4 and 5 of kLOO and kDS.

Table 4: kLOO examples: The input is a vector y for the polytope and unit group norm ball, and a matrix Y for the spechedron and unit nuclear norm ball.

Name	k best direction and output	kLOO cost
Polytope	k extreme points v_is with k smallest $\langle v, y\rangle$ among all extreme points v	See Table 3
Unit group norm ball	k groups $v_1, \ldots, v_k \in \mathcal{G}$ of the largest ℓ_2 norm of y	$\mathcal{O}((\sum_{i=1}^k \lvert v_i\rvert) + k \log k)$
Spectral simplex	bottom k eigenvector v_is of Y, output $V = [v_1, \ldots, v_k]$	Computing bottom k eigenvectors
Unit nuclear norm Ball	top k left, right singular vectors (u_i, v_i) of Y, output $U = [u_1, \ldots, u_k]$, $V = [v_1, \ldots, v_k]$.	Computing top k singular vectors

Table 5: k direction search examples. We present the parametrization of the vector x or matrix X in the second column. The kDS optimization problem is to minimize $f(x)$ or $f(X)$ over the parametrization. The input is a vector w or a matrix W in Ω and another of the form output by kLOO.

Name	Parametrization of x or X	Parameter variable	Parameter constraint (p.r.)	Main cost of proj to p.r.
Polytope	$\eta w + \sum_{i=1}^k \lambda_i v_i$	$(\eta, \lambda) \in \mathbb{R}^{k+1}$	$(\eta, \lambda) \in \Delta^{k+1}$	$\mathcal{O}(k \log(k))$
Unit Group norm ball	$\eta w + \lambda^{v_1,\ldots,v_k}$	$(\eta, \lambda^{v_1,\ldots,v_k}) \in \mathbb{R}^{1+\sum_{i=1}^k \lvert v_i\rvert}$	$\eta + \lVert\lambda^{v_1,\ldots,v_k}\rVert_{\mathcal{G}} \le 1$	$\mathcal{O}(k \log(k))+$ $\mathcal{O}(\sum_{i=1}^k \lvert v_i\rvert)$
Spectrahedron	$\eta W + VSV^\top$	$(\eta, S) \in \mathbb{R} \times \mathbb{S}^k$	$\eta \ge 0, S \succeq 0$ $\eta + \mathbf{tr}(S) = 1$	a full EVD of a k^2 matrix
Unit nuclear norm Ball	$\eta W + USV^\top$	$(\eta, S) \in \mathbb{R}^{1+k^2}$	$\eta \ge 0, \eta + \lVert S\rVert_{\text{nuc}} \le 1$	a full SVD of a k^2 matrix

A.6. Projection Step in APG for kDS of group norm ball

Here we described the projection procedure in kDS for group norm ball when the base norm is ℓ_2 norm. Suppose we want to solve the projection problem:

$$\text{minimize} \quad \lVert(\eta_0, \lambda_0^{v_1,\ldots,v_k}) - (\eta, \lambda^{v_1,\ldots,v_k})\rVert_2 \quad \text{subject to} \quad \eta + \lVert\lambda^{v_1,\ldots,v_k}\rVert_{\mathcal{G}} \le 1,\ \eta \ge 0. \tag{4}$$

We denote the optimal solution as $\eta^\star, (\lambda_0^{v_1,\ldots,v_k})^\star$.

Since $\lambda^{v_1,\ldots,v_k}$ is only supported on $\{v_i\}_{i=1}^k$, we can consider it as a vector in $\mathbb{R}^{v_1+\cdots+v_k}$ and $\lVert\lambda^{v_1,\ldots,v_k}\rVert_{\mathcal{G}} = \sum_{i=1}^k \lVert\lambda_{v_i}^{v_1,\ldots,v_k}\rVert_2$. The procedure for projection is as follows:

1. First compute the $(\eta^\star, a^\star)$ that solves

$$\text{minimize} \quad \lVert(\eta, a) - (\eta_0, [\lVert\lambda_{0,v_i}^{v_1,\ldots,v_k}\rVert]_{i=1}^k)\rVert_2, \quad \text{subject to} \quad (\eta, a) \in \mathbb{R}_+^{k+1}, \eta + \sum_{i=1}^k a_i \le 1.$$

Here $\mathbb{R}^{k+1}$ is the nonnegative orthant.

2. Next, for each v_i, we compute $(\lambda^{v_1,\dots,v_k})^\star_{v_i}$ by solving

$$(\lambda^{v_1,\dots,v_k})^\star_{v_i} = \operatorname*{arg\,min}_{\|\lambda^{v_1,\dots,v_k}_{v_i}\| \le a^\star_i} \|\lambda^{v_1,\dots,v_k}_{0,v_i} - \lambda^{v_1,\dots,v_k}_{v_i}\|_2.$$

The first step requires a projection to the convex hull of simplex and 0 and can be done in time $\mathcal{O}(k \log k)$. The second step requires projection to ℓ_2 norm ball which is a simple scaling. The correctness can be verified by decomposing each $\lambda^{v_1,\dots,v_k}_{v_i} = \alpha_i w_i$ where $\alpha_i \ge 0$ and w_i has ℓ_2 norm 1. For general ℓ_p norm, one has to find a root of a monotone function can be solved by bisection [35].

A.7. Discussion on the norm of group norm ball

For Theorems 2, the results holds for any arbitrary norm. The positive gap in Lemma 3 also holds for an arbitrary norm. However, the authors have not been able to verify whether strict complementarity implies quadratic growth for norms other than the ℓ_2 norm.

A.8. Projection Step in APG for kDS of spechedron, and nuclear norm ball

We consider how to compute the projection step of kDS for the spectrahedron and nuclear norm ball.

Spectrahedron We want to find $(\eta^\star, S^\star)$ that solves

$$\text{minimize} \quad \|(\eta, S) - (\eta_0, S_0)\|_2, \quad \text{subject to} \quad S \in \mathbb{S}^k_+,\ \eta \ge 0,\ \mathbf{tr}(S) + \eta = 1.$$

The procedures are as follows:

1. Compute the eigenvalue decomposition of $S_0 = V \Lambda_0 V^\top$, where $\Lambda_0 \in \mathbb{S}^k_+$ is a diagonal matrix with diagonal $\boldsymbol{\lambda}_0 = (\lambda_1, \dots, \lambda_k)$.
2. Compute $(\eta^\star, \boldsymbol{\lambda}^\star) = \arg\min_{(\eta,\boldsymbol{\lambda}) \in \Delta^{k+1}} \|(\eta_0, \boldsymbol{\lambda}_0) - (\eta, \boldsymbol{\lambda})\|_2$.
3. Form $S^\star = V \mathrm{diag}(\boldsymbol{\lambda}^\star) V^\top$. Here $\mathrm{diag}(\lambda)$ forms a diagonal matrix with the vector λ on the diagonal.

The main computational step is the eigenvalue decomposition which requires $\mathcal{O}(k^3)$ time. The correctness of the procedure can be verified as in [1, Lemma 3.1] and [17, Lemma 6].

Unit nuclear ball We want to find $(\eta^\star, S^\star)$ that solves

$$\text{minimize} \quad \|(\eta, S) - (\eta_0, S_0)\|_2, \quad \text{subject to} \quad \eta + \|S\|_{\text{nuc}} \le 1,\ \eta \ge 0.$$

The procedures are as follows:

1. Compute the singular value decomposition of $S_0 = U \Lambda_0 V^\top$, where $\Lambda_0 \in \mathbb{S}^k_+$ is a diagonal matrix with diagonal $\boldsymbol{\lambda}_0 = (\lambda_1, \dots, \lambda_k)$.
2. Compute $(\eta^\star, \boldsymbol{\lambda}^\star) = \arg\min_{(\eta,\boldsymbol{\lambda}) \in \Delta^{k+1}} \|(\eta_0, \boldsymbol{\lambda}_0) - (\eta, \boldsymbol{\lambda})\|_2$.
3. Form $S^\star = U \mathrm{diag}(\boldsymbol{\lambda}^\star) V^\top$. Here $\mathrm{diag}(\lambda)$ forms a diagonal matrix with the vector λ on the diagonal.

The main computational step is the singular value decomposition which requires $\mathcal{O}(k^3)$ time. The correctness of the procedure can be verified as in [1, Lemma 3.1] and [17, Lemma 6].

Appendix B. Examples, lemmas, tables, and Proofs for Section 3

B.1. Further discussion on strict complementarity

1. Traditionally, the boundary location condition $x \in \partial\Omega$ is not included in the definition of strict complementarity. We include this condition for two reasons: first, the extra location condition excludes the trivial case that the dual solution of (1) is 0, and $x_\star$ in the interior of Ω, in which case FW can be proved to converges linearly [18]; second, as we shall see in Example B.1, such assumption ensures the robustness of the sparsity of $x_\star$.

2. Strict complementarity (without the boundary location condition) holds generically: more precisely, it holds for almost all c in our optimization problem (1), $\min_{x\in\Omega} g(\mathcal{A}x) + \langle c, x\rangle$, [10, Corollary 3.5].

Example B.1 *Consider the problem*

$$\min_{x\in\alpha\Delta^n} \frac{1}{2}\|x - e_1 - \frac{1}{n}\mathbf{1}\|^2.$$

Here $\mathbf{1}$ *is the all one vector and* $\alpha > 0$*. If we set* $\alpha = 1$*, then* $x_\star = e_1$ *and the gradient* $\nabla f(x_\star) = \frac{1}{n}\mathbf{1}$*. Hence we see that strict complementarity does not hold, using Lemma 3. In this case, even though* $x_\star = e_1$ *is sparse for* $\alpha = 1$*, the solution is no longer sparse when* α *is slightly larger than* 1*. Hence, we see a perturbation to the constraint can cause instability of the rank when strict complementarity fails.*

B.2. Lemmas and tables for strict complementarity

In this section, we show that the gap quantity defined in Definition **??** is indeed positive when strict complementarity holds. We then present a table of summarizing the notations $\mathcal{F}(x_\star)$, $\mathcal{F}^c(x_\star)$, and the gap δ.

Here, for group norm ball, we consider the general norm denoted as $\|\cdot\|$ which is not necessarily the Euclidean ℓ_2 norm. The dual norm of $\|\cdot\|$ is defined as $\|x\|_* = \max_{\|y\|\le 1}\langle y, x\rangle$. We note here the group norm ball is assumed to have radius one.

Lemma 3 *When* Ω *is a polytope, group norm ball, spechedron, and nuclear norm ball, if strict complementarity holds for Problem* (1)*, then the gap* δ *is positive. Moreover, we can characterize the gradient at the solution and the size of the gap in each case:*

- *Polytope: order the vertices* $v \in \Omega$ *according to the inner products* $\langle \nabla f(x_\star), v\rangle$ *in ascending order as* $v_1, \ldots, v_{r_\star}, \ldots, v_l$ *where* l *is the total number of vertices. Then* $\langle \nabla f(x_\star), v_i\rangle$*,* $i = 1, \ldots, r_\star$ *are all equal and the gap* δ *is* $\delta = \langle \nabla f(x_\star), v_{r_\star+1}\rangle - \langle \nabla f(x_\star), v_{r_\star}\rangle$*.*
- *Group norm ball for arbitrary base norm: order vectors* $[\nabla f(x_\star)]_g$*,* $g \in \mathcal{G}$ *according to their dual norm in descending order as* $[\nabla f(x_\star)]_{g_1}, \ldots, [\nabla f(x_\star)]_{g_{|\mathcal{G}|}}$*. Then* $\|[\nabla f]_{g_i}\|_*$*,* $i = 1, \ldots, r_\star$ *are all equal, and the gap* δ *is* $\delta = \|[\nabla f(x_\star)]_{g_{r_\star}}\|_* - \|[\nabla f(x_\star)]_{g_{r_\star+1}}\|_*$*.*
- *Spectrahedron: The smallest* $r_\star$ *eigenvalues of* $\nabla f(X_\star)$ *are all equal and* $\delta = \lambda_{n-r_\star}(\nabla f(X_\star)) - \lambda_{n-r_\star+1}(\nabla f(X_\star))$
- *Nuclear norm ball: The largest* $r_\star$ *singular values of* $\nabla f(X_\star)$ *are all equal and* $\delta = \sigma_{r_\star}(\nabla f(X_\star)) - \sigma_{r_\star+1}(\nabla f(X_\star))$*.*

Proof Let us first consider the polytope case.

Polytope. Since the constraint set is a polytope and $x_\star \in \partial\Omega$, we know the smallest face $\mathcal{F}(x_\star)$ containing x is proper and admits a face-defining inequality $\langle a, x\rangle \le b$ for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. That is, $\mathcal{F}(x_\star) = \{x \mid \langle a, x\rangle = b\} \cap \Omega$ and for every $x \in \Omega$, $\langle a, x\rangle \le b$. In particular, this implies that (1) for any vertex v that is not in $\mathcal{F}(x_\star)$, $\langle a, v\rangle < b$, and (2) $\langle a, x_\star\rangle = b$.

Let us now characterize the normal cone $N_\Omega(x_\star)$. Let $\mathcal{V}$ be the set of vertices in Ω. Since Ω is bounded, we know that every point in Ω is a convex combination of the vertices. Hence $N_\Omega(x_\star)$ is the set of solutions g to the following linear system:

$$\langle g, v\rangle \le \langle g, x_\star\rangle, \quad \text{for all} \quad v \in \mathcal{V}. \tag{5}$$

Since $\mathcal{F}(x_\star)$ is the smallest face containing $x_\star$, we know that $x_\star \in \text{relint}(\mathcal{F}(x_\star))$, and so the description of normal cone $N_\Omega(x_\star)$ in (5) reduces to

$$\langle g, v_1\rangle = \langle g, x_\star\rangle, \quad \text{for all} \quad v_1 \in \mathcal{F}(x_\star), \tag{6}$$
$$\langle g, v_2\rangle \le \langle g, x_\star\rangle, \quad \text{for all} \quad v_2 \text{ being vertices of } \mathcal{F}^c(x_\star). \tag{7}$$

Note that the vector a in the face-defining inequality satisfies (6) and satisfies (7) with strict inequality as we just argued. Hence, the relative interior of $N_\Omega(x_\star)$ consists of those vectors g that satisfy (6) and satisfy (7) with a strict inequality. As $-\nabla f(x_\star) \in \text{relint}(N_\Omega(x_\star))$, we know by the previous argument that $-\nabla f(x_\star)$ satisfies (7) with strict inequality, which is exactly the condition $\delta > 0$. We arrive at the formula for δ by noting that $\langle \nabla f(x_\star), v\rangle = \langle \nabla f(x_\star), x_\star\rangle$ for every $v \in \mathcal{F}(x_\star)$ due to (6).

Group norm ball. Again, recall we here define the group norm ball using any general norm $\|\cdot\|$. The normal cone at $x_\star$ for unit group norm ball is defined as

$$N_\Omega(x_\star) = \{y \mid \langle y, x\rangle \le \langle y, x_\star\rangle, \text{ for all } \sum_{g\in\mathcal{G}} \|x_g\| \le 1\}.$$

Standard convex calculus reveals the following properties:

1. The normal cone is a linear multiple of the subdifferential for $x_\star \in \partial\Omega$: $N_\Omega(x_\star) = \{y \mid y \in \lambda\partial\|x_\star\|_\mathcal{G},\ \lambda \ge 0\}$.
2. The product rule applies to $\partial\|x_\star\|_\mathcal{G}$ as $\mathcal{G}$ forms a partition: $\partial\|x_\star\|_\mathcal{G} = \prod_{g\in\mathcal{G}} \partial\|(x_\star)_g\|$.
3. Any vector in the subdifferential of a group g in the support of the solution has norm 1: for every $g \in \mathcal{F}(x_\star)$ and every $y_g \in \partial\|(x_\star)_g\|$, $\|y_g\|_* = 1$, and $\langle y_g, (x_\star)_g\rangle = \|(x_\star)_g\|$.
4. The subdifferential for groups g not in the support is a unit dual norm ball: for every $g \in \mathcal{F}^c(x_\star)$, $\partial\|(x_\star)_g\| = \mathbf{B}_{\|\cdot\|_*} := \{y_g \in \mathbb{R}^{|g|} \mid \|y_g\|_* \le 1\}$.

The above properties reveal that the normal cone is the set

$$N_\Omega(x_\star) = \Big\{y \mid y \in \lambda \left(\prod_{g\in\mathcal{F}(x_\star)} \partial\|(x_\star)_g\| \times \prod_{g\in\mathcal{G}-\mathcal{F}(x_\star)} \mathbf{B}_{\|\cdot\|_*}\right), \lambda \ge 0\Big\}, \tag{8}$$

where for every $g \in \mathcal{F}(x_\star)$ and every $y_g \in \partial\|(x_\star)_g\|$, $\|y_g\|_* = 1$. Hence, we know that the relative interior of $N_\Omega(x_\star)$ is simply

$$\text{relint}\left(N_\Omega(x_\star)\right) = \Big\{y \mid y \in \lambda \left(\prod_{g\in\mathcal{F}(x_\star)} \text{relint}\left(\partial\|(x_\star)_g\|\right) \times \prod_{g\in\mathcal{G}-\mathcal{F}(x_\star)} \text{relint}\left(\mathbf{B}_{\|\cdot\|_*}\right)\right), \lambda > 0\Big\}, \tag{9}$$

where for every $g \in \mathcal{F}(x_\star)$, and every $y_g \in \text{relint}\left(\partial\|(x_\star)_g\|\right)$, $\|y_g\|_* = 1$, and for every $g \in \mathcal{G} - \mathcal{F}(x_\star)$, and every $y_g \in \text{relint}\left(\mathbf{B}_{\|\cdot\|_*}\right)$, $\|y_g\|_* < 1$. Because of the strict inequality of λ in (9), and strict inequality for $\|y_g\|_* < 1$ for $y_g \in \text{relint}\left(\mathbf{B}_{\|\cdot\|_*}\right)$, we see that

$$\|[\nabla f(x_\star)]_{g_1}\|_* = \cdots = \|[\nabla f(x_\star)]_{g_{r_\star}}\|_*, \text{ and } \|[\nabla f(x_\star)]_{g_{r_\star}}\|_* - \|[\nabla f(x_\star)]_{g_{r_\star+1}}\|_* > 0 \tag{10}$$

as $-\nabla f(x_\star) \in \text{relint}\left(N_\Omega(x_\star)\right)$. Using the condition that for every $g \in \mathcal{F}(x_\star)$ and every $y_g \in \partial\|(x_\star)_g\|$, $\|y_g\|_* = 1$, and $\langle y_g, (x_\star)_g\rangle = \|(x_\star)_g\|$, we know $\langle -\nabla f(x_\star), x_\star\rangle = \|[\nabla f(x_\star)]_{g_{r_\star}}\|_*$. Furthermore, using generalized Cauchy-Schwarz, it can be proved that $\min_{x\in\mathcal{F}^c(x_\star)}\langle\nabla f(x_\star), x\rangle = -\|[\nabla f(x_\star)]_{g_{r_\star+1}}\|_*$. Hence, combining the two equalities with (10), we see that $\delta > 0$ and arrive at the stated formula for δ.

Spectrahedron. We first note that $X_\star \in \partial\Omega$ and $\mathbf{tr}(X) = 1$ imply that $1 \le r_\star < n$. To compute the normal cone, we can apply the sum rule of subdifferentials to

$$\chi(\{X \in \mathbb{S}^n \mid \mathbf{tr}(X) = 1\}) + \chi(X \succeq 0),$$

where χ is the characteristic function, which takes value 0 for elements belonging to the set and $+\infty$ otherwise) of $\{X \in \mathbb{S}^n \mid \mathbf{tr}(X) = 1\}$ and $\mathbb{S}^n_+$ and reach

$$N_\Omega(X_\star) = \{sI \mid s \in \mathbb{R}\} + \{-Z \mid Z \succeq 0, \text{range}(Z) \subset \text{nullspace}(X_\star)\}. \tag{11}$$

We note that sum rule can be applied here because $\frac{1}{n}I$ belongs to the interior of both sets.

Next, using the sum rule of relative interior to (11) , we find that

$$\text{relint}(N_\Omega(X_\star)) = \{sI \mid s \in \mathbb{R}\} + \{-Z \mid Z \succeq 0, \text{range}(Z) = \text{nullspace}(X_\star)\}. \tag{12}$$

Or equivalently,

$$\text{relint}(N_\Omega(X_\star)) = \{sI \mid s \in \mathbb{R}\} + \{-Z \mid Z \succeq 0, \text{nullspace}(Z) = \text{range}(X_\star)\}. \tag{13}$$

Denote the eigenspace corresponding to the smallest $r_\star$ values of $\nabla f(X_\star)$ as $\mathbf{EV}_{r_\star}(\nabla f(X_\star))$. Using (13), it is immediate that $-\nabla f(X_\star) \in \text{relint}(N_\Omega)$ means that

$$\mathbf{EV}_{r_\star}(\nabla f(X_\star)) = \text{range}(X_\star).$$

. Moreover, from (13) and $-\nabla f(X_\star) \in \text{relint}(N_\Omega)$, we also have

$$\lambda_{n-r_\star+1}(\nabla f(X_\star)) - \lambda_{n-r_\star}(\nabla f(X_\star)) > 0, \quad \text{and} \quad \langle\nabla f(X_\star), X_\star\rangle = \lambda_{n-i}(\nabla f(X_\star)),\ i = 1, \ldots, r_\star. \tag{14}$$

Combining (14) and the well-known fact that

$$\min_{X\in\Omega,\, \text{range}(X)\perp\mathbf{EV}_{r_\star}(\nabla f(X_\star))=\text{range}(X_\star)} \langle\nabla f(X_\star), X\rangle = \lambda_{n-r_\star+1}(\nabla f(X_\star)),$$

we see that δ is indeed positive, and the formula for δ holds.

Nuclear norm ball. We first note that $X_\star \in \partial\Omega$ imply that $1 < r_\star < \min(n_1, n_2)$, and $\|X_\star\|_{\text{nuc}} = 1$. Let the singular value decomposition of $X_\star$ as $X_\star = U\Sigma V$ with $U \in \mathbb{R}^{n_1 \times r_\star}$ and $V \in \mathbb{R}^{n_2 \times r_\star}$. The normal cone of the unit nuclear norm ball is

$$N_\Omega(X_\star) = \{Y \mid Y = \lambda Z,\ Z = UV^\top + W,\ W^\top U = 0,\ WV = 0,\ \|W\|_{\text{op}} \le 1 \text{ and } \lambda \ge 0\}. \quad (15)$$

Hence, the relative interior is

$$N_\Omega(X_\star) = \{Y \mid Y = \lambda Z,\ Z = UV^\top + W,\ W^\top U = 0, WV = 0, \|W\|_{\text{op}} < 1 \text{ and } \lambda > 0\}. \quad (16)$$

Since $-\nabla f(X_\star) \in \text{relint}(N_\Omega(X_\star))$, we know immediately that

$$\sigma_{r_\star}(\nabla f(X_\star)) - \sigma_{r_\star+1}(\nabla f(X_\star)) > 0, \quad (17)$$

$$(18)$$

the top $r_\star$ left and right singular vectors of $\nabla f(X_\star)$ are just the columns of $-U$ and V, and $\langle \nabla f(X_\star), X_\star\rangle = -\sigma_i(\nabla f(X_\star))$ for $i = 1, \ldots, r_\star$. Combining pieces and the standard fact that

$$\min_{\text{range}(X)\perp\text{range}(X_\star)=\text{range}(U),\, \|X\|_{\text{nuc}}\le 1} \langle \nabla f(X_\star), X\rangle = -\sigma_{r_\star+1}(\nabla f(X_\star)),$$

we see the gap δ is indeed positive and the formula is correct. ■

A table of the notions $\mathcal{F}^c(x_\star), \mathcal{F}^c(x_\star)$, and the formula of gap δ is shown as Table 6.

Table 6: For several constraint sets Ω, this table describes the support set $\mathcal{F}(x_\star)$, its complementary set $\mathcal{F}^c$, and the gap δ. Recall the gap $\delta = \min\{\langle u, \nabla f(x_\star)\rangle - \langle x_\star, \nabla f(x_\star)\rangle \mid u \in \mathcal{F}^c(x_\star) \subset \Omega\}$ and admits a specific formula as described in Lemma 3. We denote the gradient at $x_\star$ as $\nabla_\star$. For a polytope, we order the vertices $v \in \Omega$ according to their inner products $\langle \nabla f(x_\star), v\rangle$ in ascending order as $v_1, \ldots, v_{r_\star}, \ldots, v_l$, where l is the number of vertices. For the group norm ball, we order vectors $[\nabla f(x_\star)]_g$, $g \in \mathcal{G}$ according to their ℓ_2 norm in descending order as $[\nabla f(x_\star)]_{g_1}, \ldots, [\nabla f(x_\star)]_{g_{|\mathcal{G}|}}$.

Constraint Ω	$\mathcal{F}(x_\star)$	$\mathcal{F}^c$	δ formula
polytope	smallest face containing $x_\star$	convex hull of all the vertices not in $\mathcal{F}(x_\star)$	$\langle \nabla f(x_\star), v_{r_\star+1}\rangle - \langle \nabla f(x_\star), v_{r_\star}\rangle$
group norm ball	$\{g \in \mathcal{G} \mid (x_\star)_g \neq 0\}$	$\{x \mid x_g = 0, \forall g \in \mathcal{F}(x_\star)\}$	$\|[\nabla_\star]_{g_{r_\star}}\|_2 - \|[\nabla_\star]_{g_{r_\star+1}}\|_2$
Spectrahedron	range$(X_\star)$	$\{X \in [\text{range}(X_\star)]^\perp\} \cap \mathcal{SP}^n$	$\lambda_{n-r_\star}(\nabla_\star) - \lambda_{n-r_\star+1}(\nabla_\star)$
Nuclear norm ball	range$(X_\star)$	$\{X \in [\text{range}(X_\star)]^\perp\} \cap \mathbf{B}_{\|\cdot\|_{\text{nuc}}}$	$\sigma_{r_\star}(\nabla_\star) - \sigma_{r_\star+1}(\nabla_\star)$

B.3. Quadratic growth under strict complementarity

This section develops that quadratic growth does hold under strict complementarity and the condition g in (1) is strongly convex.

Theorem 4 *Suppose Problem* (1)*,* $\min_{x\in\Omega} g(\mathcal{A}x) + \langle c, x\rangle$*, satisfies that* g *is strongly convex and the constraint set* Ω *is one of the four sets (i) polytope, (ii) unit group norm ball, (iii) spechedron, and (vi) unit nuclear norm ball. Further suppose that strict complementarity holds. Then quadratic growth holds for Problem* (1) *as well.*

We will use the machinery developed in [38] for the case of the group norm. We define a few notions and notations for later convenience. We define the projection to Ω as $\mathcal{P}_\Omega(x) := \arg\min_{v\in\Omega} \|x-v\|_2$. The difference of iterates for projected gradient with step size t is defined as $\mathcal{G}_t(x) := \frac{1}{t}(x - \mathcal{P}_\Omega(x - \nabla f(x)))$. Note that $\mathcal{G}_t(x) = 0$ implies $x = x_\star$. Finally, for an arbitrary set $\mathcal{S}$, we define the distance of $x \in \mathbb{R}^n$ to it as $\text{dist}(x, \mathcal{S}) := \inf_{v\in\mathcal{S}} \|x - v\|_2$.

Proof The proof for the polytope appears in [4, Lemma 2.5]. The proof of the Spectrahedron appears in [9, Theorem 6]. Here, we address the case of the group norm ball and Nuclear norm ball. Let us first consider the case of the group norm ball with the ℓ_2 norm.

Unit group norm ball. Using [11, Corollary 3.6], we know that if the error bound condition holds for some $t, \gamma > 0$ then the quadratic growth condition holds with some parameter γ'. The error bound condition with parameter $t, \gamma, \epsilon > 0$ means that for all $x \in \Omega$ and $\|x - x_\star\|_2 \le \epsilon$, the following the inequality holds: [6]

$$\|x - x_\star\|_2 \le \gamma \|\mathcal{G}_t(x)\|_2. \tag{19}$$

Define $\bar{y} = \mathcal{A}(x_\star)$ and $\bar{h} = \nabla f(x_\star)$. Now using [38, Corollary 1 and Theorem 2], we need only verify the following two conditions to establish (19):

1. *Bounded linear regularity:* The two sets $\Gamma_f(\bar{y}) := \{x \in \mathbf{E} \mid \bar{y} = \mathcal{A}(x)\}$ and $\Gamma_\Omega(\bar{h}) := \{x \in \mathbf{E} \mid -\bar{h} \in N_\Omega(x)\}$ satisfy that for every bounded set B, there exists a constant κ such that
$$\text{dist}(x, \Gamma_f(\bar{y}) \cap \Gamma_\Omega(\bar{h})) \le \kappa \max\{\text{dist}(x, \Gamma_f(\bar{y})), \text{dist}(x, \Gamma_\Omega(\bar{h}))\}, \quad \text{for all} \quad x \in \Omega.$$
2. *Metric subregularity:* there exists $\kappa, \epsilon > 0$ such that for all x with $\|x - x_\star\|_2 \le \epsilon$,
$$\text{dist}(x, \Gamma_\Omega(\bar{h})) \le \kappa \text{dist}(-\bar{h}, N_\Omega(x)). \tag{20}$$

Let us first verify bounded linear regularity. First, the subdifferential of the Euclidean norm $\|\cdot\|_2$ is

$$\partial\|x\|_2 = \begin{cases} \frac{x}{\|x\|_2} & x \neq 0, \\ \mathbf{B}_{\|\cdot\|_2} & x = 0. \end{cases}.$$

From the characterization (9), we know that $\bar{h} = \nabla f(x_\star)$ is nonzero due to strict complementarity, and hence any $x \in \Gamma(\bar{h})$ must satisfy $x \in \partial\Omega$. Following the derivation of the normal cone in (8), we have for any $x \in \partial\Omega$,

$$N_\Omega(x) = \left\{y \mid y \in \lambda \left(\prod_{g\in\mathcal{F}(x)} \partial\|x_g\|_2 \times \prod_{g\in\mathcal{G}-\mathcal{F}(x)} \mathbf{B}_{\|\cdot\|_2} \right),\ \lambda \ge 0 \right\}. \tag{21}$$

6. The error bound condition considered in [11, Corollary 3.6] actually require the bound (19) to hold for all x in the intersection of Ω and a sublevel set of f. Note there is a difference between a sublevel set and a neighborhood of $x_\star$. Because the continuity of f and the compactness of Ω, we know the two are actually equivalent (with a different choice of γ). Moreover, the quadratic growth considered there is only required to hold for x in Ω and a sublevel set of f. Again, this is not a problem as Ω is compact and f is continuous.

Here the support set $\mathcal{F}(x)$ is the set of groups in the support of x. Let us pick a $i^\star \in \mathcal{F}(x_\star)$. For each $i \in \mathcal{G}$, define the vector $\tilde{h}_i = \frac{-\bar{h}_i}{\|h_{i^\star}\|_2} \in \mathbb{R}^{|v_i|}$. Recall from (10), we have $\|\bar{h}_i\|_2$ all equal for $i \in \mathcal{F}(x_\star)$. For each $i \in \mathcal{F}(x_\star)$, define $\tilde{h}^i \in \mathbb{R}^n$ so that it is only supported on group i with vector value $\tilde{h}_i$ and is 0 elsewhere. Again, from (10) and Lemma 3, we have $\|\bar{h}^i\|_2$ all equal for $i \in \mathcal{F}(x_\star)$, and is larger than those i not in $\mathcal{F}(x_\star)$. To remember the notation, we use $\tilde{h}^i$, upper index i, to mean vector in $\mathbb{R}^n$. We use the notation $\tilde{h}_i$, lower index i, to mean the shorter vector in $\mathbb{R}^{|v_i|}$.

Combining the facts about $\tilde{h}_i$, the formula (21), the formula of $\|\cdot\|_2$, and $x \in \partial\Omega$, we find that actually

$$\Gamma_\Omega(\bar{h}) = \{x \mid \sum_{i \in \mathcal{F}(x_\star)} \alpha_i \tilde{h}^i, \alpha_i \in \Delta^{|\mathcal{F}(x_\star)|}\},$$

which is a convex polyhedral. Because $\Gamma_f(\bar{y})$ and $\Gamma_\Omega(\bar{h})$ are both convex polyhedral, we know from [3, Corollary 3] that bounded linear regularity holds.

We verify metrical subregularity now. Note that from previous calculation of $\Gamma_\Omega(\bar{h})$, we know

$$\text{dist}(x, \Gamma_\Omega(\bar{h}))^2 = \min_{\alpha_i \in \Delta^{|\mathcal{F}(x_\star)|}} \sum_{i \in \mathcal{F}(x_\star)} \|x_i - \alpha_i \tilde{h}_i\|_2^2 + \sum_{i \notin \mathcal{F}(x_\star)} \|x_i\|_2^2.$$

By choosing ϵ sufficiently small, say $\epsilon < \epsilon_0$, we have $\mathcal{F}(x) \supset \mathcal{F}(x_\star)$. The quantity, $\text{dist}(\bar{h}, N_\Omega(x))$, on the RHS of (20) for all x within an ϵ neighborhood of the solution $x_\star$ satisfies that

$$\text{dist}^2(\bar{h}, N_\Omega(x)) = \begin{cases} +\infty, & x \notin \Omega, \\ \|\bar{h}\|_2^2, & x \in \text{int}(\Omega), \end{cases}$$

where $\text{int}(\Omega)$ is the interior of Ω. For $x \in \partial\Omega$, $\text{dist}^2(\bar{h}, N_\Omega(x))$ satisfies that

$$\text{dist}^2(\bar{h}, N_\Omega(x)) = \|h_{i^\star}\|_2^2 \min_{\lambda \geq 0, v_i \in \mathbf{B}_{\|\cdot\|_2}} \sum_{i \in \mathcal{F}(x)} \|\tilde{h}_i - \frac{\lambda x_i}{\|x_i\|_2}\|_2^2 + \sum_{i \notin \mathcal{F}(x)} \|\tilde{h}_i - \lambda v_i\|_2^2.$$

The case of $x \notin \Omega$ is trivial. The case of $x \in \text{int}(\Omega)$ can be proved by choosing a large enough κ, say $\kappa > K_0$, as $\text{dist}(x, \Gamma_\Omega(\bar{h}))^2$ is upper bounded for any $x \in \Omega$, and $\text{dist}^2(\bar{h}, N_\Omega(x))$ in this case is fixed. We are left with the most challenging case $x \in \partial\Omega$, where the normal cone is non-trivial. First, we upper bound $\text{dist}(x, \Gamma_\Omega(\bar{h}))^2$ by choosing $\alpha_i = \|x_i\|_2$. The numbers α_i sum to one because $x \in \partial\Omega$. In this case, $\text{dist}(x, \Gamma_\Omega(\bar{h}))^2$ satisfies the bound

$$\begin{aligned} \text{dist}(x, \Gamma_\Omega(\bar{h}))^2 &\leq \sum_{i \in \mathcal{F}(x_\star)} \|x_i - \|x_i\|_2 \tilde{h}_i\|_2 + \sum_{i \notin \mathcal{F}(x_\star)} \|x_i\|_2^2 \\ &\overset{(a)}{=} \sum_{i \in \mathcal{F}(x_\star)} \|x_i\|_2 \|\tilde{h}_i - \frac{x_i}{\|x_i\|_2}\|_2^2 + \sum_{i \in (\mathcal{F}(x) - \mathcal{F}(x_\star))} \|x_i\|_2^2, \end{aligned} \tag{22}$$

where step (a) is due to $\mathcal{F}(x) \supset \mathcal{F}(x_\star)$ by our choice of small enough ϵ. We next lower bound $\text{dist}^2(\bar{h}, N_\Omega(x))$ by ignoring the term not in $\mathcal{F}(x)$:

$$\text{dist}^2(\bar{h}, N_\Omega(x)) \geq \|g_{i^\star}\|_2^2 \min_{\lambda \geq 0} \sum_{i \in \mathcal{F}(x)} \|\tilde{h}_i - \frac{\lambda x_i}{\|x_i\|_2}\|_2^2$$

Now if $\mathcal{F}(x) = \mathcal{F}(x_\star)$, then it is tempting to set $\lambda = 1$ above and compare the inequality with (22) to claim victory. This does not work directly due to the minimization over λ and the fact $\mathcal{F}(x) \supset \mathcal{F}(x_\star)$.

Let $\lambda_\star = \arg\min_{\lambda \geq 0} \sum_{i \in \mathcal{F}(x)} \|\tilde{h}_i - \frac{\lambda x_i}{\|x_i\|_2}\|_2^2$. In this case, we have an explicit formula of $\lambda^\star$:

$$\lambda_\star = \max\left\{0, \frac{\sum_{i \in \mathcal{F}(x)} \langle \tilde{h}_i, \frac{x_i}{\|x_i\|_2}\rangle}{|\mathcal{F}(x)|}\right\}.$$

If $\lambda_\star = 0$, then we can simply pick some $\kappa > K_0$ as done in the case of $x \in \mathrm{int}(\Omega)$. So we assume $\lambda_\star > 0$ in the following. Next let $\lambda_i = \arg\min_{\lambda \geq 0} \|\tilde{h}_i - \frac{\lambda x_i}{\|x_i\|_2}\|_2^2$ for each $i \in \mathcal{F}(x_\star)$. With such choice of λ_i and $\lambda^\star$, we can further lower bound $\mathrm{dist}^2(\bar{h}, N_\Omega(x))$ by splitting the terms in $\mathcal{F}(x)$ and those are not:

$$\mathrm{dist}^2(\bar{h}, N_\Omega(x)) \geq \|g_{i^\star}\|_2^2 \left(\underbrace{\sum_{i \in \mathcal{F}(x_\star)} \|\tilde{h}_i - \frac{\lambda_i x_i}{\|x_i\|_2}\|_2^2}_{R_1} + \underbrace{\sum_{i \in \mathcal{F}(x) \setminus \mathcal{F}(x_\star)} \|\tilde{h}_i - \frac{\lambda_\star x_i}{\|x_i\|_2}\|_2^2}_{R_2} \right). \tag{23}$$

We bound the two terms separately. Let us first deal with R_1. From the expression of normal cone (21) and $-\bar{h} \in N_\Omega(x_\star)$ by our assumption, we know $\tilde{h}_i = \frac{(x_\star)_i}{\|(x_\star)_i\|_2}$ for every $i \in \mathcal{F}(x_\star)$. Hence by choosing a (possibly smaller) ϵ, say $\epsilon < \epsilon_1$, we can ensure that for any x within an ϵ_1 neighborhood of the solution $x_\star$, $\langle \frac{x_i}{\|x_i\|_2}, \tilde{h}_i \rangle \geq 0$ all for $i \in \mathcal{F}(x_\star)$. Moreover, for a small enough ϵ_1, we know each $\lambda_i = \langle \frac{x_i}{\|x_i\|_2}, \tilde{h}_i \rangle$ and is very close to 1. Thus the condition of Lemma 5 is fulfilled, and we have

$$R_1 \geq \frac{1}{2} \sum_{i \in \mathcal{F}(x_\star)} \|x_i\|_2 \|\tilde{h}_i - \frac{x_i}{\|x_i\|_2}\|_2^2. \tag{24}$$

Next, to deal with R_2, let us examine the expression of $\lambda_\star = \frac{\sum_{i \in \mathcal{F}(x)} \langle \tilde{h}_i, \frac{x_i}{\|x_i\|_2}\rangle}{|\mathcal{F}(x)|}$. Recall $\langle \tilde{h}_i, \frac{x_i}{\|x_i\|_2}\rangle$ is close to 1 for small enough ϵ. Due to strict complementarity, for each $i \in \mathcal{F}(x) \setminus \mathcal{F}(x_\star)$, we know $\|\tilde{h}_i\|_2 < 1 - \delta_0$ for some $\delta_0 > 0$ that depends only on $\bar{h}$. Combining these two facts together, we know that $i' = \arg\min_{i \in \mathcal{F}(x)} \langle \tilde{h}_i, \frac{x_i}{\|x_i\|_2}\rangle$ must belong to $\mathcal{F}(x) \setminus \mathcal{F}(x_\star)$. Moreover, by choosing an even smaller ϵ, say $\epsilon < \epsilon_2$, we have $\lambda_\star \geq \delta_1 + \min_i \langle \tilde{h}_i, \frac{x_i}{\|x_i\|_2}\rangle$ for some $\delta_1 > 0$ that only depends on $\bar{h}$, δ, and ϵ_2. We can now lower bound R_2 as follows:

$$\begin{aligned} R_2 &\geq \|\tilde{h}_{i'} - (\delta_1 + \langle \tilde{h}_i, \frac{x_i}{\|x_i\|_2}\rangle) \frac{x_{i'}}{\|x_{i'}\|_2}\|_2^2 \\ &= \|\tilde{h}_i - \langle \tilde{h}_i, \frac{x_i}{\|x_i\|_2}\rangle \frac{x_{i'}}{\|x_{i'}\|_2}\|_2^2 + \delta_1^2 + 2\delta \underbrace{\langle \tilde{h}_i - \langle \tilde{h}_i, \frac{x_i}{\|x_i\|_2}\rangle \frac{x_{i'}}{\|x_{i'}\|_2}, \frac{x_{i'}}{\|x_{i'}\|_2}\rangle}_{=0} \\ &\geq \delta_1^2. \end{aligned} \tag{25}$$

Combining the bounds (24) and (25) on R_1 and R_2, we found that

$$\begin{aligned}\operatorname{dist}^2(\bar{h}, N_\Omega(x)) &\geq \frac{\|g_{i^\star}\|_2^2}{2} \sum_{i\in\mathcal{F}(x_\star)} \|\tilde{h}_i - \frac{x_i}{\|x_i\|_2}\|_2^2 + \|g_{i^\star}\|_2^2\delta_1^2 \\ &\overset{(a)}{\geq} \frac{\|g_{i^\star}\|_2^2}{2} \sum_{i\in\mathcal{F}(x_\star)} \|\tilde{h}_i - \frac{x_i}{\|x_i\|_2}\|_2^2 + \|g_{i^\star}\|_2^2\delta_1^2 \sum_{i\in\mathcal{F}(x)\setminus\mathcal{F}(x_\star)} \|x_i\|_2^2\end{aligned} \tag{26}$$

Here, for the step (a), we use $\|x_i\|_2 \leq 1$ as $x \in \partial\Omega$. Hence, by taking $\epsilon = \min(\epsilon_1, \epsilon_2)$ and $\kappa = \max\{K_0, \|g_{i^\star}\|_2^2\delta_1, \frac{\|g_{i^\star}\|_2^2}{2}\}$, and comparing (26) with (22), a bound on $\operatorname{dist}(x, \Gamma_\Omega(\bar{h}))^2$, we see that metric subregularity is satisfied and our proof for unit group norm ball is complete.

Finally, we consider the unit nuclear norm ball.

Unit nuclear norm ball. Let us first illustrate the main idea. We shall utilize the quadratic growth result proved in [9, Theorem 6] for spectrahedron. To transfer our setting to spectrahedron, we use a dilation argument with its relating lemmas [8, Lemma 3] and [26, lemma 1]. We now spell out all the details.

Let $\tilde{n} = n_1 + n_2$. For any $\tilde{X} \in \mathbb{S}^{\tilde{n}}$ as $\lambda_1(\tilde{X}) \geq \cdots \geq \lambda_{\tilde{n}}(\tilde{X})$. Also, for any $X \in \mathbf{B}_{\|\cdot\|_{\text{nuc}}}$, denote is SVD as $X = U_X\Sigma_X V_X^\top$ where $U_X \in \mathbb{R}^{n_1\times r_X}$, $V_X \in \mathbb{R}^{n_2\times r_X}$, and $r_X = \operatorname{rank}(X)$. Define the dilation $X^\sharp \in \mathbb{S}^{\tilde{n}}$ of a $X \in \mathbb{R}^{n_1\times n_2}$ as

$$X^\sharp = \frac{1}{2}\begin{bmatrix} X_1 & X \\ X^\top & X_2 \end{bmatrix}, \tag{27}$$

where the $X_1 = U_X(\Sigma_X + \xi_X I)U_X^\top$, and $X_2 = V_X(\Sigma_X + \xi_X I)V_X^\top$. The number $\xi_X \geq 0$ is chosen so that $\tilde{X}$ has trace 1. Note that $X^\sharp$ is positive semidefinite as $X^\sharp = \frac{1}{2}\begin{bmatrix} U_X \\ V_X \end{bmatrix}\Sigma_X[U_X^\top V_X^\top] + \frac{\xi_X}{2}\begin{bmatrix} U_X U_X^\top & 0 \\ 0 & V_X V_X^\top \end{bmatrix}$. Also define the linear map $\mathcal{B} \in \mathbb{S}^{\tilde{n}} \to \mathbb{R}^{n_1\times n_2}$ such that for any $\tilde{Y} = \begin{bmatrix} Y_1 & Y \\ Y^\top & Y_2 \end{bmatrix} \in \mathbb{S}^{\tilde{n}}$ with $Y_1 \in \mathbb{S}^{n_1}$, and $Y_2 \in \mathbb{S}^{n_2}$,

$$\mathcal{B}(\tilde{Y}) := 2Y.$$

Consider the problem

$$\begin{aligned}\text{minimize} \quad & \tilde{f}(\tilde{X}) := f(\mathcal{B}(\tilde{X})) = g(\mathcal{A}(\mathcal{B}\tilde{X})) + \langle \mathcal{B}^* c, \tilde{X}\rangle \\ \text{subject to} \quad & \tilde{X} = 1 \quad \tilde{X} \succeq 0.\end{aligned} \tag{28}$$

We claim that it satisfies strict complementarity and its solution $\tilde{X}_\star$ is unique and is simply $X_\star^\sharp$. Suppose the claim is proved for the moment, then using [9, Theorem 6], we know there is some $\tilde{\gamma} > 0$, such that for all $\tilde{X} \in \mathcal{SP}^{\tilde{n}}$, we have

$$\tilde{f}(\tilde{X}) - \tilde{f}(X_\star^\sharp) \geq \gamma\|\tilde{X} - X_\star^\sharp\|_{\text{F}}.$$

Hence, for any $X \in \Omega$, by construction of $\tilde{f}$, we have

$$f(X) - f(X_\star) = \tilde{f}(X^\sharp) - \tilde{f}(X_\star^\sharp) \geq \gamma\|X^\sharp - X_\star^\sharp\|_{\text{F}}^2 \geq \frac{\gamma}{2}\|X - X_\star\|_{\text{F}}^2.$$

This proves quadratic growth.

We now verify our claim that $X_\star^\sharp$ is the unique solution to (28) and $X_\star^\sharp \in \partial\mathcal{SP}^{\tilde{n}}$ with $\nabla\tilde{f}(X_\star^\sharp) \in \partial N_{\mathcal{SP}^{\tilde{n}}}(X_\star^\sharp)$. First, consider feasibility and whether $X_\star^\sharp \in \partial\mathcal{SP}^{\tilde{n}}$. The condition $X_\star \in \partial\Omega$ implies that $1 \leq r_\star < \min(n_1, n_2)$ and $\|X_\star\|_{\text{nuc}} = 1$. Hence we do have $\mathbf{tr}(X^\sharp) = 1$ and $X_\star^\sharp \in \partial\mathcal{SP}^{\tilde{n}}$ as $\text{rank}(X_\star^\sharp) = \text{rank}(X_\star) = r_\star < n_1 + n_2$. Next, consider optimality. Given any $\tilde{X} \in \mathcal{SP}^{\tilde{n}}$ with $\tilde{X} = \frac{1}{2}\begin{bmatrix} X_1 & X \\ X^\top & X_2 \end{bmatrix}$, by [26, Lemma 1], we have

$$\|X\|_{\text{nuc}} \leq 1. \tag{29}$$

To see $X_\star^\sharp$ is optimal for (28), note that

$$f(\mathcal{B}(\tilde{X})) = f(X) \overset{(a)}{\geq} f(X_\star) = f(\mathcal{B}(X_\star^\sharp)),$$

where step (a) is due to optimality of $X_\star$ in (1) and X is feasible as just argued. Thirdly, we argue that $X_\star^\sharp$ is a unique solution to (28). For any optimal solution $\tilde{X}_\star = \frac{1}{2}\begin{bmatrix} X_1^\star & X_0 \\ X_0^\top & X_2^\star \end{bmatrix}$ of (28), we have X_0 is optimal to (1) as

$$f(X_0) = f(\mathcal{B}(\tilde{X}_\star)) \overset{(a)}{=} f(\mathcal{B}(X_\star^\sharp)) = f(X_\star),$$

where step (a) is because $X_\star^\sharp$ is optimal to (28). Hence due to uniqueness of $X_\star$, we know $X_0 = X_\star$. Because $\|X_\star\|_{\text{nuc}} = 1$, using [8, Lemma 3], we know in fact $X_\star^\sharp = \tilde{X}_\star$ and uniqueness of solution to (28) is proved. Finally, we verify strict complementarity that $\nabla\tilde{f}(X_\star^\sharp) \in \text{relint}\left(N_{\mathcal{SP}^{\tilde{n}}}(X_\star^\sharp)\right)$. Recall from (13), that we need to show

$$-\nabla\tilde{f}(X_\star^\sharp) \in \text{relint}(N_{\mathcal{SP}^{\tilde{n}}}(\tilde{X}_\star)) = \{sI \mid s \in \mathbb{R}\} + \{-\tilde{Z} \mid \tilde{Z} \succeq 0, \text{range}(\tilde{Z}) = \text{nullspace}(X_\star^\sharp)\}. \tag{30}$$

Using the definition of $X_\star^\sharp$, we know

$$\nabla\tilde{f}(X_\star^\sharp) = \begin{bmatrix} 0 & \nabla f(X_\star) \\ \nabla f(X_\star)^\top & 0 \end{bmatrix}.$$

Recall from Lemma 3, we have $\sigma_1(\nabla f(X_\star)) = \cdots = \sigma_{r_\star}(\nabla f(X_\star)) = \delta + \sigma_{r_\star+1}(\nabla f(X_\star))$ for some gap $\delta > 0$. Hence we see that $\nabla\tilde{f}(X_\star^\sharp)$ has all its smallest $r_\star$ eigenvalues equal as $-\sigma_{r_\star}(\nabla f(X_\star))$ and the gap between its $r_\star$-th smallest eigenvalue and the $r_\star + 1$-th eigenvalue is simply $\delta > 0$. Moreover, let the singular value decomposition of $X_\star$ as $X_\star = U_\star \Sigma V_\star^\top$ with $U_\star \in \mathbb{R}^{n_1 \times r_\star}$ and $V_\star \in \mathbb{R}^{n_2 \times r_\star}$. From the description of normal cone of nuclear norm ball in (16), we know $U_\star, -V_\star$ are the matrices formed by the top $r_\star$ left and right vectors of $\nabla f(X_\star)$. Hence, the bottom $r_\star$ eigenvector of $\nabla f(\tilde{X}_\star)$ is simply $\frac{1}{\sqrt{2}}\begin{bmatrix} U_\star \\ V_\star. \end{bmatrix}$. Since $\text{range}(X_\star^\sharp) = \text{range}(\begin{bmatrix} U_\star \\ V_\star. \end{bmatrix})$, we may take $s = \sigma_1(\nabla f(X_\star))$ and $\tilde{Z} = \sigma_1(\nabla f(X_\star))I + \nabla\tilde{f}(X_\star^\sharp)$. Using the eigengap condition on $\nabla\tilde{f}(X_\star^\sharp)$, we see $\text{range}(\tilde{Z}) = \text{nullspace}(X_\star^\sharp)$ and our claim is proved. ■

B.3.1. ADDITIONAL LEMMA FOR QUADRATIC GROWTH

We establish the following lemma for the proof of unit group norm ball.

Lemma 5 *For any two $x, y \in \mathbb{R}^d$ with ℓ_2 norm one, and $a := \langle x, y\rangle \geq 0$, we have*

$$2 \min_{\lambda \geq 0} \|x - \lambda y\|_2^2 \geq \|x - y\|_2^2.$$

Proof Simple calculus reveals that the optimal solution $\lambda^\star$ of the LHS of the inequality is $\lambda^\star = a \geq 0$. We know $a \in [0, 1]$ due to Cauchy-Schwarz and our assumption on a. Direct calculation of the difference yields

$$\begin{aligned} 2 \min_{\lambda \geq 0} \|x - \lambda y\|_2^2 - \|x - y\|_2^2 &= 2 + 2a^2 - 4a^2 - 2 + 2a \\ &= -2a^2 + 2a \geq 0, \end{aligned} \tag{31}$$

where the last line is due to $a \in [0, 1]$. ■

B.4. Proofs of Theorem ?? for group norm ball

Proof The inequality (**??**) follows from the proof of convergence of FW as in [25] by noting that the vector $v_t = \arg\min_{v \in \Omega} \langle \nabla f(x_t), v\rangle$ is feasible for the kDS minimization problem. ■

Proof The proof follows from the intuition that once x_t is close to $x_\star$, the set $\mathcal{F}(x_\star)$ can be identified using $\nabla f(x_t)$. The fact that $\delta > 0$ is shown in Lemma 3. Let us now consider Algorithm 2 whose constraint set Ω is a polytope. Denote the objective gap as $h_t = f(x_t) - f(x_\star)$. Using quadratic growth in the following step (a), and Theorem 2 in the following second step (b), and the choice in the following step (c), the iterate x_t with $t \geq T$ satisfies that

$$\|x_t - x_\star\| \overset{(a)}{\leq} \sqrt{\frac{1}{\gamma} h_t} \overset{(b)}{\leq} \sqrt{\frac{L_f D^2}{\gamma T}} \overset{(c)}{\leq} \frac{\delta}{2 L_f D}. \tag{32}$$

Next, for any $t \geq T$, we have that for any vertex v in $\mathcal{F}(x_\star)$, and any vertices u in $\mathcal{F}^c(x_\star)$,

$$\begin{aligned} \langle \nabla f(x_t), v\rangle - \langle \nabla f(x_t), u\rangle &= \langle \nabla f(x_\star), v - u\rangle + \langle \nabla f(x_t) - \nabla f(x_\star), v - u\rangle \\ &\overset{(a)}{\leq} -\delta + \langle \nabla f(x_t) - \nabla f(x_\star), v - u\rangle \overset{(b)}{\leq} -\frac{\delta}{2}. \end{aligned} \tag{33}$$

Here in step (a), we use the definition of δ in (**??**) and $\langle x_\star, \nabla f(x_\star)\rangle = \langle v, \nabla f(x_\star)\rangle$ using the optimality condition for Problem (1) and $\mathcal{F}(x_\star)$ being the smallest face containing $x_\star$. In step (b), we use the bound in (32) , Lipschitz continuity of $\nabla f(x)$, and $\|v - u\| \leq D$.

Thus, the kLOO step will produce all the vertices in $\mathcal{F}(x_\star)$ as $k \geq r_\star$ after $t \geq T$, and so $x_\star$ is a feasible and optimal solution of the optimization problem in the $k$direction search step. Hence, Algorithm 2 finds the optimal solution $x_\star$ within $T + 1$ many steps. The case for unit group norm ball can be similarly analyzed and we defer the detail to Section B.4 in the appendix. ■

Proof Let us now consider Algorithm 2 whose constraint set Ω is a unit group norm ball with *arbitrary* base norm $\|\cdot\|$. Using quadratic growth (a), Theorem 2 in the second step (b), and the choice of T in the following step (c), the iterate x_t with $t \geq T$ satisfies that

$$\|x_t - x_\star\| \overset{(a)}{\leq} \sqrt{\frac{1}{\gamma} h_t} \overset{(b)}{\leq} \sqrt{\frac{L_f D^2}{\gamma T}} \overset{(c)}{\leq} \frac{\delta}{2L_f D}. \tag{34}$$

Next recall the definition of $\mathcal{F}(x_\star)$ implies $(x_\star)_{g_\star} \neq 0$ for any $g_\star \in \mathcal{F}(x_\star)$. The optimality conditions and $\|x_\star\|_{\mathcal{G}} = 1$ (due to $x_\star \in \partial\Omega$) implies that for every $g_\star \in \mathcal{F}(x_\star)$,

$$\langle [\nabla f(x_\star)]_{g_\star}, [x_\star]_{g_\star} \rangle = \|[\nabla f(x_\star)]_{g_\star}\|_* \|[x_\star]_{g_\star}\|, \quad \text{and} \quad \langle \nabla f(x_\star), x_\star \rangle = \|[\nabla f(x_\star)]_{g_\star}\|_*.$$

For any $g_\star \in \mathcal{F}(x_\star)$, define a vector $x_\star^{g_\star} \in \Omega$ as $x_\star^{g_\star} := \begin{cases} \left(\frac{[x_\star]_{g_\star}}{\|[x_\star]_{g_{r_\star}}\|}\right)_i, & i \in g_\star, \\ 0, & i \notin g_\star. \end{cases}$ So $x_\star^{g_\star} \in \Omega$ is an extended vector of the normalized vector $\frac{[x_\star]_{g_\star}}{\|[x_\star]_{g_{r_\star}}\|}$. Combining this definition with previous two equalities, we see

$$\langle [\nabla f(x_\star)]_{g_\star}, x_\star^{g_\star} \rangle = \langle \nabla f(x_\star), x_\star \rangle. \tag{35}$$

Now, for any $t \geq T$, we have for any group $g_\star$ in $\mathcal{F}(x_\star)$, and any vector $v \in \Omega$ that is in $\mathcal{F}^c(x_\star)$,

$$\begin{aligned} \langle \nabla f(x_t), x_\star^{g_\star} \rangle - \langle \nabla f(x_t), v \rangle &= \langle \nabla f(x_\star), x_\star^{g_\star} - v \rangle + \langle \nabla f(x_t) - \nabla f(x_\star), v - u \rangle \\ &\overset{(a)}{\leq} -\delta + \langle \nabla f(x_t) - \nabla f(x_\star), v - u \rangle \overset{(b)}{\leq} -\frac{\delta}{2}. \end{aligned} \tag{36}$$

Here in step (a), we use the definition of δ in (**??**) and (35). In step (b), we use the bound in (32) , Lipschitz continuity of $\nabla f(x)$, and $\|v - u\| \leq D$.

Thus, the kLOO step will produce all the groups in $\mathcal{F}(x_\star)$ as $k \geq r_\star$ after $t \geq T$, and so $x_\star$ is a feasible and optimal solution of the optimization problem in the k direction search step. Hence Algorithm 2 finds the optimal solution $x_\star$ within $T + 1$ steps. ■

B.5. Proofs of Theorem ??

We state one lemma that is critical to our proof of linear convergence. It is proved in Section B.5.1.

Lemma 6 *Given $Y \in \mathbb{R}^{n_1 \times n_2}$ with $\sigma_r(Y) - \sigma_{r+1}(Y) = \delta > 0$. Denote the matrices formed by the top r left and right singular vectors of Y as $U \in \mathbb{R}^{n_1 \times r}$, $V \in \mathbb{R}^{n_2 \times r}$ respectively. Then for any $X \in \mathbb{R}^{n_1 \times n_2}$ with $\|X\|_{\text{nuc}} = 1$,*

there is an $S \in \mathbb{R}^{r \times r}$ with $\|S\|_{\text{nuc}} = 1$ such that

$$\langle X - USV^\top, Y \rangle \geq \frac{\delta}{2} \|X - USV^\top\|_{\text{F}}^2.$$

Equipped with this lemma, let us now prove Theorem **??**.

Proof [Proof of Theorem **??**] The case of the spectrahedron is proved in [9, Theorem 3] by using the eigengap formula in Lemma 3 and [9, Section 2.2 "relation with the eigengap assumption"]. Here, we need only to address the case of the unit nuclear norm. The proof that we present here for the case

of the nuclear norm ball is quite similar. For notation convenience, define the set $\mathcal{N}_{r_\star,t} = \{USV^\top \mid U, V$ are matrices formed by top $r_\star$ left and right singular vectors of $\nabla f(X_t)$ and $\|S\|_{\text{nuc}} \le 1\}$.

Using the Lipschitz smoothness of f, we have for any $t \ge 1$, $\eta \in [0,1]$, and any $W \in \mathcal{N}_{r_\star,t}$:

$$\begin{aligned} f(X_{t+1}) \le & f(X_t) + (1-\eta)\langle W - X_t, \nabla f(X_t)\rangle \\ & + \frac{(1-\eta)^2 L_f}{2}\|W - X_t\|_{\text{F}}^2. \end{aligned} \tag{37}$$

For $t \ge T$, we find that $\|X_t - X_\star\|_{\text{F}} \le \frac{\delta}{6\sqrt{2}L_f}$, and

$$\begin{aligned} & \sigma_{r_\star}(\nabla f(X_t)) - \sigma_{r_\star+1}(\nabla f(X_t)) \\ = & \underbrace{\sigma_{r_\star}(\nabla f(X_\star)) - \sigma_{r_\star+1}(\nabla f(X_\star))}_{\overset{(a)}{=} -\delta} + \underbrace{(\sigma_{r_\star}(\nabla f(X_t)) - \sigma_{r_\star+1}(\nabla f(X_\star)))}_{\overset{(b)}{\le} \frac{1}{3}\delta} \\ & + \underbrace{(\lambda_{n-r_\star+1}(\nabla f(X_\star)) - \lambda_{n-r_\star+1}(\nabla f(X_t)))}_{\overset{(c)}{\le} \frac{1}{3}\delta} \\ \le & -\frac{1}{3}\delta. \end{aligned}$$

Here in step (a), we use the singular value gap formula of δ in Lemma 3. Step (b) and (c) are due to Weyl's inequality, the Lipschitz continutity of ∇f, and the inequality $\|X_t - X_\star\|_{\text{F}} \le \frac{\delta}{6\sqrt{2}L_f}$.

Now we subtract the inequality (37) both sides by $f(X_\star)$, and denote $h_t = f(X_t) - f(X_\star)$ for each t to arrive at

$$\begin{aligned} h_{t+1} \le & h_t + (1-\eta)\underbrace{\langle W - X_t, \nabla f(X_t)\rangle}_{R_1} \\ & + \frac{(1-\eta)^2 L_f}{2}\underbrace{\|W - X_t\|_{\text{F}}^2}_{R_2}. \end{aligned} \tag{38}$$

Using Lemma 6, the inequality (38), and the assumption $X_\star \in \partial\Omega$, we can choose $W \in \mathcal{N}_{r_\star,t}$ such that

$$\langle W - X_\star, \nabla f(X_t)\rangle \le -\frac{\delta}{6}\|X_\star - W\|_{\text{F}}^2. \tag{39}$$

Let us now analyze the term $R_1 = \langle W - X_t, \nabla f(X_t)\rangle$ using (39) and convexity of f:

$$\begin{aligned} R_1 = & \langle W - X_t, \nabla f(X_t)\rangle \\ = & \langle W - X_\star, \nabla f(X_t)\rangle + \langle X_\star - X_t, \nabla f(X_t)\rangle \\ \le & -\frac{\delta}{6}\|X_\star - W\|_{\text{F}}^2 - h_t. \end{aligned}$$

The term $R_2 = \|X_t - W\|_{\text{F}}^2$ can be bounded by

$$\begin{aligned} R_2 = \|X_t - W\|_{\text{F}}^2 & \overset{(a)}{\le} 2\left(\|X_t - X_\star\|_{\text{F}}^2 + \|X_\star - W\|_{\text{F}}^2\right) \\ & \overset{(b)}{\le} \frac{2}{\gamma}h_t + 2\|X_\star - W\|_{\text{F}}^2, \end{aligned}$$

where we use the triangle inequality and the basic inequality $(a+b)^2 \leq 2a^2 + 2b^2$ in step (a), and the quadratic growth condition in step (b).

Now combining (38), and the bounds of R_1 and R_2, we reach that there is a $W \in \mathcal{N}_{r_\star,t}$ such that for any $\xi = 1-\eta \in [0,1]$, we have

$$\begin{aligned} h_{t+1} \leq & h_t + \xi\left(-\frac{\delta}{6}\|X_\star - W\|_{\mathrm{F}}^2 - h_t\right) + \frac{\xi^2 L_f}{2}\left(\frac{2}{\gamma}h_t + 2\|X_\star - W\|_{\mathrm{F}}^2\right) \\ = & \left(1 - \xi + \frac{\xi^2 L_f}{\gamma}\right)h_t + \left(\xi^2 L_f - \frac{\xi\delta}{6}\right)\|X_\star - W\|_{\mathrm{F}}^2. \end{aligned}$$

A detailed calculation and choice of ξ below yields the factor $1 - \min\{\frac{\gamma}{4L_f}, \frac{\delta}{12L_f}\}$ in the theorem.

We need to choose $\xi \in [0,1]$ so that $1 - \xi + \frac{\xi^2 L_f}{\gamma}$ is minimized while keeping $\xi^2 L_f - \frac{\xi\delta}{6} \leq 0$. For $\xi^2 L_f - \frac{\xi\delta}{6} \leq 0$, we need $\xi \leq \frac{\delta}{6L_f}$. The function $q(\xi) = 1 - \xi + \frac{\xi^2 L_f}{\gamma}$ is decreasing for $\xi \leq \frac{\gamma}{2L_f}$ and increasing for $\xi \geq \frac{\gamma}{2L_f}$. If $\frac{\gamma}{2L_f} \leq \frac{\delta}{6L_f}$, then we can pick $\xi = \frac{\gamma}{2L_f}$, and $q(\xi) = 1 - \frac{\gamma}{4L_f}$. If $\frac{\gamma}{2L_f} \geq \frac{\delta}{6L_f} \implies \frac{\delta}{\gamma} \leq 3$, then we can pick $\xi = \frac{\delta}{6L_f}$, and $q(\xi) = 1 - \frac{\delta}{6L_f} + \frac{\delta^2}{36\gamma L_f} = 1 + \frac{\delta}{6L_f}\left(\frac{\delta}{6\gamma} - 1\right) \leq 1 - \frac{\delta}{12L_f}$. ■

B.5.1. ADDITIONAL LEMMAS FOR THE PROOF OF THEOREM ??

Here we give a proof of Lemma 6.

Proof [Proof of Lemma 6] We utilize the result in [9, Lemma 5]: given any $\tilde{Y} \in \mathbb{S}^n$ with eigenvalues $\lambda_n(\tilde{Y}) \leq \cdots \leq \lambda_{n-r+1}(\tilde{Y}) \leq \lambda_{n-r}(\tilde{Y}) - \delta' \leq \cdots \leq \lambda_1(\tilde{Y}) - \delta'$ for some $\delta' > 0$. Denote the matrices by the bottom r eigenvectors of $\tilde{Y}$ as $\tilde{V} \in \mathbb{R}^{n_2 \times r}$ respectively. Then for any $\tilde{X} \in \mathbb{S}_+^n$ with $\mathbf{tr}(\tilde{X}) = 1$, there is an $S \in \mathbb{R}_+^{r\times r}$ with $\mathbf{tr}(S) = 1$ such that

$$\langle \tilde{X} - \tilde{V}S\tilde{V}^\top, \tilde{Y}\rangle \geq \frac{\delta'}{2}\|\tilde{X} - \tilde{V}S\tilde{V}^\top\|_{\mathrm{F}}^2. \tag{40}$$

To utilize this result, we consider the dilation of the matrices X and Y:

$$\tilde{X} := \frac{1}{2}\begin{bmatrix} X_1 & X \\ X^\top & X_2 \end{bmatrix}, \quad \text{and} \quad \tilde{Y} := \begin{bmatrix} 0 & Y \\ Y^\top & 0 \end{bmatrix}. \tag{41}$$

Here the matrices $X_1 = U_X\Sigma_X U_X$, $X_2 = V_X\Sigma_X V_X^\top$ where $U_X\Sigma_X V_X$ is the SVD of X and the number $r_X = \mathrm{rank}(X)$. Since $\tilde{X} = \begin{bmatrix} U_X \\ V_X \end{bmatrix}\Sigma_X[U_X^\top \; V_X^\top]$, the matrix $\tilde{X} \in \mathbb{S}_+^{n_1+n_2} \succeq 0$. The trace of $\tilde{X}$ is $\mathbf{tr}(\tilde{X}) = 1$ as $\|X\|_{\mathrm{nuc}} = 1$. Note that the bottom $r+1$ eigenvalues of $\tilde{Y}$ is simply $-\sigma_1(Y), \ldots, -\sigma_r(Y), -\sigma_{r+1}(Y)$, and the matrix $\tilde{V} \in \mathbb{R}^{(n_1+n_2)r}$ defined below is formed by the matrix eigenvectors corresponds the smallest r eigenvalues:

$$\tilde{V} := \frac{1}{\sqrt{2}}\begin{bmatrix} U \\ -V \end{bmatrix}. \tag{42}$$

Using [9, Lemma 5], we can find a matrix $S \in \mathbb{S}_+^r$ with $\mathbf{tr}(S) = 1$ such that (40) holds. Writing the equation in block form reveals that

$$\langle X - U(-S)V^\top, Y\rangle = \langle \tilde{X} - \tilde{V}S\tilde{V}^\top, \tilde{Y}\rangle \geq \frac{\delta}{2}\|\tilde{X} - \tilde{V}S\tilde{V}^\top\|_{\mathrm{F}}^2 \overset{(a)}{\geq} \frac{\delta}{2}\|X - U(-S)V^\top\|_{\mathrm{F}}^2, \tag{43}$$

where the last step is due to Lemma 7. Note that the matrix $U(-S)V^\top$ is the matrix we seek as $\| - S\|_{\text{nuc}} = \mathbf{tr}(S) = 1$. Hence the proof is completed. ■

Lemma 7 *Suppose two matrices $X, Y \in \mathbb{R}^{n_1 \times n_2}$, and $X = U_1 S_1 V_1^\top$ and $Y = U_2 S_2 V_2^\top$ for some unitary U_i, V_i that for some integers r_1, r_2, they satisfy $U_i \in \mathbb{R}^{n_1 \times r_i}$, $i = 1, 2$ and $V_i \in \mathbb{R}^{n_2 \times r_i}$, $i = 1, 2$. The matrices $S_i \in \mathbb{S}_+^{r_i}$ are positive semidefinite. Then*

$$\|U_1 S_1 U_1^\top - U_2 S_2 U_2^\top\|_{\text{F}}^2 + \|V_1 S_1 V_1^\top - V_2 S_2 V_2^\top\|_{\text{F}}^2 \geq 2\|U_1 S_1 V_1^\top - U_2 S_2 V_2^\top\|_{\text{F}}^2.$$

Proof This result follows by direct computation. Consider the difference $\|U_1 S_1 U_1^\top - U_2 S_2 U_2^\top\|_{\text{F}}^2 + \|V_1 S_1 V_1^\top - V_2 S_2 V_2^\top\|_{\text{F}}^2 - 2\|U_1 S_1 V_1^\top - U_2 S_2 V_2^\top\|_{\text{F}}^2$. Expanding the square and using the orthogonal invariance of the Frobenius norm, we find that

$$\begin{aligned}
&\|U_1 S_1 U_1^\top - U_2 S_2 U_2^\top\|_{\text{F}}^2 + \|V_1 S_1 V_1^\top - V_2 S_2 V_2^\top\|_{\text{F}}^2 - 2\|U_1 S_1 V_1^\top - U_2 S_2 V_2^\top\|_{\text{F}}^2 \\
=&2\mathbf{tr}(S_1 U_1^\top U_2 S_2 (U_2^\top U_1 - V_2^\top V_1)) + 2\mathbf{tr}(S_1 (V_1^\top V_2 - U_1^\top U_2) S_2 V_2^\top V_1) \\
\overset{(a)}{=}&2\mathbf{tr}(S_1 U_1^\top U_2 S_2 (U_2^\top U_1 - V_2^\top V_1)) + 2\mathbf{tr}(V_1^\top V_2 S_2 (V_2^\top V_1 - U_2^\top U_1) S_1) \\
\overset{(b)}{=}&2\mathbf{tr}(S_1 (U_1^\top U_2 - V_1^\top V_2) S_2 (U_2^\top U_1 - V_2^\top V_1)),
\end{aligned}$$

where step (a) is due to the fact that $\mathbf{tr}(A) = \mathbf{tr}(A^\top)$ and step (b) is due to the cyclic property of trace. By factorizing $S_i = S_i^{\frac{1}{2}}$ for $i = 1, 2$ and the cyclic property of trace again, we find that

$$\begin{aligned}
&\|U_1 S_1 U_1^\top - U_2 S_2 U_2^\top\|_{\text{F}}^2 + \|V_1 S_1 V_1^\top - V_2 S_2 V_2^\top\|_{\text{F}}^2 - 2\|U_1 S_1 V_1^\top - U_2 S_2 V_2^\top\|_{\text{F}}^2 \\
=&\mathbf{tr}(S_1^{\frac{1}{2}} (U_1^\top U_2 - V_1^\top V_2) S_2^{\frac{1}{2}} S_2^{\frac{1}{2}} (U_2^\top U_1 - V_2^\top V_1) S_1^{\frac{1}{2}}) \\
=&\|S_2^{\frac{1}{2}} (U_2^\top U_1 - V_2^\top V_1) S_1^{\frac{1}{2}}\|_{\text{F}}^2 \geq 0.
\end{aligned}$$

Hence the lemma is proved. ■

Appendix C. Generalization to atomic sets

We consider how to generalize the ideas in this paper to other atomic sets.

Suppose we have a set of atoms $\mathcal{A}$ that each $a \in \mathcal{A}$ is in the same finite dimensional space $\mathbf{E}$. There might be infinitely many atoms in $\mathcal{A}$. We assume for any $a_i \in \mathcal{A}$, the norm of $\|a_i\|_2$ is uniformly bounded by $\frac{D}{2}$ for some $D \geq 0$. Then we consider $\Omega = \mathbf{conv}(\mathcal{A})$. For the sets considered in the paper, the sets $\mathcal{A}$ are

1. Polytope: $\mathcal{A}$ has finite cardinality.
2. Group norm ball: All vectors in $\mathbf{B}_{\|\|_{\mathcal{G}}}$ that are only supported in one group, and different vectors can have different support group.
3. Spectrahedron: rank one positive semidefinite matrices with trace 1.
4. Unit nuclear norm: rank one matrices with nuclear norm less than or equal to one.

Suppose the optimal solution $x_\star \in \Omega$ admits a minimal representation using atoms $a_i^\star$: $r'_\star$ is the smallest number such that $x_\star \in \mathbf{conv}([a_i]_{i=1}^{r'_\star})$. To further define kLOO and kDS, we must first examine the optimality condition in this case. The optimality condition reveals that

$$\langle \nabla f(x_\star), x_\star \rangle = \langle \nabla f(x_\star), a_i^\star \rangle \leq \langle \nabla f(x_\star), a \rangle, \quad \forall \quad a \in \mathcal{A}. \tag{44}$$

The core idea enabling the proof of algorithms is that there should be a separation in the above inequality. Denote the dual vector as $\nabla_\star := \nabla f(x_\star)$. Also define the exposed face defined by $\nabla f(x_\star)$ as $\mathcal{F}(\nabla_\star) = \Omega \cap \{x \mid \langle x, \nabla_\star \rangle = \langle x_\star, \nabla_\star \rangle\}$. Since $\mathbf{conv}([a_i]_{i=1}^{r'_\star}) \neq \mathcal{F}(\nabla_\star)$ in general, this means that we might need to compute more than $r'_\star$ vectors in finding the vs that minimize $\langle v, \nabla_\star \rangle$. This observation explains why we need did not define $r_\star$ to be the number smallest number of vertices in polytope and why we needed to redefine kDS for the spectrahedron and nuclear norm ball.

The key of enabling faster convergence is the gap condition. This means that we need to decompose Ω into two convex sets Ω_1 and Ω_2 such that $\Omega_1 \cap \Omega_2 = \emptyset$, $\mathbf{conv}(\Omega_1, \Omega_2) = \Omega$, $\Omega_1 \supset \mathcal{F}(\nabla_\star)$, and there is a gap $\delta > 0$,

$$\langle \nabla f(x_\star), x_\star \rangle = \langle \nabla f(x_\star), a_i^\star \rangle \leq \langle \nabla f(x_\star), a \rangle - \delta, \quad \forall \quad a \in \mathbf{conv}(\Omega_2). \tag{45}$$

For the Ω considered in the paper, we have

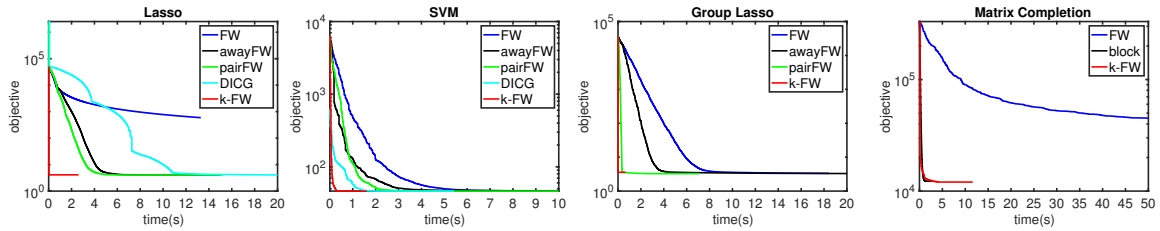
1. Polytope: $\Omega_1 = \mathcal{F}(x_\star)$, and $\Omega_2 = \mathcal{F}^c(x_\star)$
2. Group norm ball: $\Omega_1 = \{x \in \Omega, x \text{ supported in } \mathcal{F}(x_\star)\}$, and $\Omega_2 = \mathcal{F}^c(x_\star)$.
3. Spectrahedron or nuclear norm ball: $\Omega_1 = \{X \mid \text{range}(X) \in \mathcal{F}(X_\star)\}$, and $\Omega_2 = \mathcal{F}^c(X_\star)$.

Let us now explain how to generalize the ideas in this paper to atomic sets more concretely:

1. Algorithmically, it is feasible to compute $\Omega_1 \supset \mathcal{F}(\nabla_\star)$ and optimize over it given $\nabla_\star$. The kLOO and kDS are approximate versions of Ω_1 for general $\nabla f(x)$.
2. Theoretically, in order to prove faster convergence, we need to find Ω_1 and Ω_2, that induces the gap δ. A usual candidate of Ω_1 is $\mathcal{F}(\nabla_\star)$. Linearity, orthogonality, and other properties of Ω should be considered for such choice.

Appendix D. Discussion on multiple solutions

When the problem has more than one solution, let $\mathcal{X}$ be the solution set. This set is convex and closed. We change the term $\|x - x_\star\|$ in the quadratic growth condition to $\text{dist}(x, \mathcal{X}) = \min_{x_\star \in \mathcal{X}} \|x - x_\star\|$. For strict complementarity, we remove the condition that $x_\star$ is unique and demand instead that some $x_\star \in \mathcal{X}$ satisfies the conditions listed in strict complementarity. The support set $\mathcal{F}(x_\star)$ and complementary set $\mathcal{F}^c(x_\star)$ are defined via the $x_\star$ that satisfies strict complementarity. Note that the dual vector $\nabla f(x_\star)$ is the same for every $x_\star \in \mathcal{X}$ [38, Proposition 1]. The algorithmic results, Theorem 2, **??**, and **??** hold almost without any change of the proof using the new definition of $r_\star$ and δ. The argument to establish quadratic growth via strict complementarity is more tedious and we defer it to future work.



(a) (b) (c) (d)

Figure 3: Objective against time cost

Appendix E. Numerical Experiment setting for Section 4

We detail the experiment settings of Lasso, support vector machine (SVM), group Lasso, and matrix completion problems. The compared methods include FW, away-step FW (awayFW) [22], pairwise FW (pairFW)[29], DICG [19], and blockFW [1]. All codes are written by MATLAB and performed on a MacBook Pro with Processor 2.3 GHz Intel Core i5 and Memory 8 GB 2133 MHz LPDDR3. In our k-FW, we solve the kDS by the FASTA toolbox [20, 21]: *https://github.com/tomgoldstein/fasta-matlab*. In DICG (as well as FW, awayFW, pairFW in Group Lasso and SVM), the step size is determined by backtracking line search. The ball sizes of ℓ_1 norm, group norm, and nuclear norm are set to be the ground truth respectively.

E.1. Lasso

The experiment is the same as that in [29] except that the data size in our setting is ten times of that in [29]: $A \in \mathbb{R}^{2000\times 5000}$ and $b \in \mathbb{R}^{2000}$. The large size is more reasonable for comparing the computational costs of FW, awayFW, pairFW, DICG and our k-FW. For FW, awayFW and pairFW, we use the MATLAB codes provide by [29]: `https://github.com/Simon-Lacoste-Julien/linearFW`. In DICG (as well as FW, awayFW, pairFW in Group Lasso and SVM), the step size is determined by backtracking line search.

E.2. SVM

We generate the synthetic data for two-class classification by the following model

$$X = [X_1 \;\; X_2] = [U_1V_1 + 1 \;\; U_2V_2 - 1], \quad X \leftarrow X + E,$$

where the elements of $U_1 \in \mathbb{R}^{20\times 5}$, $V_1 \in \mathbb{R}^{5\times 500}$, $U_2 \in \mathbb{R}^{20\times 5}$, and $V_2 \in \mathbb{R}^{5\times 500}$ are drawn from $\mathcal{N}(0,1)$. E consists of noise drawn from $\mathcal{N}(0, 0.1\sigma_X)$, where σ_X denotes the standard deviation of the entries of X. Thus, in X, the number of samples is 1000 and the number of features is 20. We use 80% of the data as training data to classify the remaining data. In SVM, we use a polynomial kernel $k(x, y) = (x^\top y + 1)^2$.

E.3. Group Lasso

We generate a 100×1000 matrix X whose entries are drawn from $\mathcal{N}(0, 1)$ and a 10×100 matrix W with 10 nonzero columns drawn from $\mathcal{N}(0, 1)$. Then let $Y = WX$ and set $Y \leftarrow Y + E$, where the entries of noise matrix E are drawn from $\mathcal{N}(0, 0.01\sigma_Y)$. Then we estimate W from Y and X by solving a Group Lasso problem with kFW.

E.4. Matrix Completion

We generate a low-rank matrix as $X = UV\top$, where the entries of $U \in \mathbb{R}^{500\times 5}$ and $V \in \mathbb{R}^{5\times 500}$ are drawn from $\mathcal{N}(0, 1)$. We sample 50% of the entries uniformly at random and recover the unknown entries by low-rank matrix completion.

E.5. Objective function vs running time

See Figure 3.