Two Spectral Gaps: Decentralized Optimization over Intersections of Local Convex Sets

Sarvagya Jha

Preprint. Submitted to the *Journal of Optimization Theory and Applications.*

Abstract We study decentralized minimization of an average of strongly convex, smooth local objectives over an intersection of agent-private closed convex sets, where each agent knows only its own objective and its own set and agents communicate over a gossip network. We show that the complexity is controlled by a single geometric scalar, which we call the constraint gap: the reciprocal square of the averaged linear regularity modulus of the collection of sets. For affine sets it coincides with a spectral gap of the averaged projection operator. We prove that a positive constraint gap forces the strong conical hull intersection property, bound the perturbation of the solution map, and show that for affine local sets the dual condition number is at most the product of the objective condition number, the gossip condition number and the reciprocal constraint gap, with an explicit family attaining that bound exactly. We give a Chebyshev-accelerated dual method whose oracle and communication complexities are square roots of these quantities, and validate the theory numerically. A companion paper shows that the same operator governs a known primal method and proves a matching lower bound in the constraint gap.

Keywords Decentralized optimization · Linear regularity · Friedrichs angle · Convex feasibility · Oracle complexity · Gossip algorithms

Mathematics Subject Classification (2000) 90C25 · 90C60 · 65K05 · 47N10 · 93A16

Sarvagya Jha
Graduate Research Assistant, University of Massachusetts Amherst
Amherst, MA 01003, USA
ORCID: 0009-0004-5078-8227
E-mail: jhasarvagya@gmail.com

1 Introduction

Let $G=(V,E)$ be a connected undirected graph on n agents. Agent i holds a private convex function $f_i:\mathbb{R}^d\to\mathbb{R}$ and a private closed convex set $X_i\subseteq\mathbb{R}^d$, and may exchange vectors only with its neighbours. The agents wish to solve

$$\min_{x\in\mathbb{R}^d} f(x):=\frac{1}{n}\sum_{i=1}^n f_i(x) \qquad \text{subject to} \qquad x\in X:=\bigcap_{i=1}^n X_i. \tag{P}$$

Problem (P) models multi-agent decision making under distributed feasibility requirements: each agent contributes both a share of the cost and a share of the constraints, and no agent ever sees the whole feasible set.

1.1 The obstruction

For *unconstrained* decentralized optimization ($X_i=\mathbb{R}^d$) the complexity picture is complete: with μ-strongly convex, L-smooth f_i, Scaman et al. [29] proved matching bounds of $\Theta(\sqrt{\kappa}\log\frac{1}{\varepsilon})$ local computations and $\Theta(\sqrt{\kappa\chi}\log\frac{1}{\varepsilon})$ communication rounds, where $\kappa=L/\mu$ and χ is the gossip condition number.

The constrained problem behaves differently. Writing $r_i=\iota_{X_i}$, (P) is a composite problem with *agent-specific* nonsmooth terms, and Alghunaim et al. [1, Thm. 3, Cor. 1] show that for any method accessing the data through $J_i,\nabla J_i,r_i$ and $\operatorname{prox}_{r_i}$ there are ν-strongly convex instances, in dimension $M=O(KG/\sqrt{\nu\varepsilon})$, on which $O(G/\sqrt{\nu\varepsilon})$ or more iterations are needed; no such method is globally linearly convergent. Consistently, decentralized algorithms handling genuinely heterogeneous X_i converge sublinearly [17, 22, 26, 31, 33, 34, 37], and the one work carrying a geometric regularity constant through a complexity bound, Mai and Abed [20], places it in the constants of an $O(\log t/\sqrt{t})$ rate. The exception is Pang [24], who obtains *asymptotic* linear convergence for the decentralized *feasibility* problem (no objective) with agent-private sets, under a local nondegeneracy constraint qualification and with an unquantified contraction factor.

1.2 The constraint gap

Our starting point is

$$\sigma \;:=\; \inf_{x\in\mathbb{R}^d\setminus X} \frac{\frac{1}{n}\sum_{i=1}^n \operatorname{dist}^2(x,X_i)}{\operatorname{dist}^2(x,X)} \;\in\; [0,1], \tag{1.1}$$

the *constraint gap* of $\{X_1,\dots,X_n\}$: the largest constant for which the averaged error bound $\operatorname{dist}^2(x,X)\le\sigma^{-1}\frac{1}{n}\sum_i\operatorname{dist}^2(x,X_i)$ holds. Positivity of σ is exactly linear regularity in the sense of Bauschke and Borwein [3], a classical condition: automatic for polyhedra by Hoffman's error bound, valid under a

Slater-type interior condition, and equivalent to subtransversality of the collection [15, 16].

For *affine* sets σ has an exact spectral meaning. With $\mathcal{X} = X_1 \times \cdots \times X_n$ and $\mathcal{C}$ the consensus subspace in Pierra's product space [25],

$$\sigma \;=\; 1 - c_F(\mathcal{X}_0, \mathcal{C})^2 \;=\; 1 - \left\|\bar{P} - P_{X^0}\right\|, \qquad \bar{P} = \tfrac{1}{n}\sum\nolimits_i P_{X_i^0}, \tag{1.2}$$

where c_F is the Friedrichs angle cosine and X_i^0, X^0 denote direction subspaces, c_F being taken between the product set and the consensus subspace (Remark 3.5 warns against a competing convention). **This identification is not ours**: the operator-norm form $\|\bar{P} - P_{X^0}\| = c_F^2$ is Reich and Zalas [27, Thm. 8, Rem. 9] at $k = 1$, extended to arbitrary weights and countable families in Reich and Zalas [28, Thms. 1.6, 3.3]; the passage to the distance form (1.2) is one line, which we give in Remark 3.6. Meanwhile the qualitative transfer of linear regularity between $\{X_i\}$ and $(\prod_i X_i, \mathcal{C})$ for arbitrary sets is Dao et al. [5, Prop. 3.11(a)]. We recall it in Section 3 because σ read as a *spectral gap* is what makes the rest of the paper work. On the network side the gossip matrix W is a Laplacian-like *residual* operator — $\langle x, \mathbf{W}x\rangle$ measures disagreement and vanishes exactly on the consensus subspace — so it is the mixing operator $I - \mathbf{W}$ that drives iterates towards consensus, contracting by $1 - \gamma$. On the constraint side the residual operator is $I - \bar{P}$, whose quadratic form is the average squared infeasibility, and the corresponding averaging operator — the simultaneous projection $Q = \frac{1}{n}\sum_i P_{X_i}$, whose linear part is $\bar{P}$ — drives iterates towards feasibility, contracting by $1 - \sigma$. The two mechanisms are formally identical (Table 2), and the complexity of (P) depends on them in the same way.

1.3 Contributions

We state at the outset that the complexity theory of Sections 5–6 assumes the local sets are *affine* (Assumption 5.1); Sections 3 and 4 hold for general closed convex sets.

1. **Optimality and stability under** $\sigma > 0$ (Section 4). $\sigma > 0$ implies the strong conical hull intersection property — this implication is Bauschke et al. [4, Thm. 3], not ours — so $\bar{x}$ solves (P) if and only if there exist $v_i \in N_{X_i}(\bar{x})$ with $\frac{1}{n}\sum_i(\nabla f_i(\bar{x}) + v_i) = 0$: the optimality certificate itself decomposes across agents. We bound the perturbation of the solution under simultaneous perturbation of the f_i and the X_i, with explicit σ-dependence.
2. **Dual conditioning, with a matching extremal family** (Section 5). For affine local sets the dual formed by dualizing consensus alone has condition number at most $\kappa\chi/\sigma$ (Theorem 5.3), via a spectral lemma for products ARA with A positive definite and R positive semidefinite (Lemma 5.2). Theorem 5.5 exhibits, for every prescribed $\kappa \geq 1$, $\chi \geq 1$ and

$\sigma \in (0, \frac{1}{2}]$, an explicit instance in $\mathbb{R}^3$ attaining $\kappa_{\text{dual}} = \kappa\chi/\sigma$ *exactly*, so none of the three factors can be removed or improved.

3. **Algorithm and upper bound** (Section 6). CANDA reaches accuracy ε with $O(\sqrt{\kappa/\sigma}\log\frac{1}{\varepsilon})$ local oracle calls and $O(\sqrt{\kappa\chi/\sigma}\log\frac{1}{\varepsilon})$ communication rounds. Setting $\sigma = 1$ recovers the optimal unconstrained rates; setting $\kappa = \chi = 1$ recovers the optimal rate for convex feasibility. Because σ is a global quantity no agent can form, Section 6.4 shows the guarantee degrades gracefully under an underestimate, discusses a heuristic power iteration and what it does *not* certify, and identifies structured cases in which a valid lower bound on σ is available in closed form.
4. **σ is intrinsic, and it is unavoidable** (companion paper [12]). Two results are established there and used here only as context. First, in the quadratic–affine setting PG-EXTRA — a primal method built with no reference to σ — reduces exactly to a second-order recursion driven by $\frac{1}{4}P_{\mathcal{X}_0}\mathbf{W}P_{\mathcal{X}_0}$, whose nonzero spectrum is that of the operator behind Theorem 5.3, and no step size makes it faster than $\sqrt{(\kappa-1)\chi/\sigma}$ on the family of Theorem 5.5. Second, in a decentralized projection black-box class every method needs $\Omega(\sigma^{-1/2}\log\frac{1}{\varepsilon})$ rounds, so the exponent $\frac{1}{2}$ in σ cannot be improved; specialized to two subspaces this gives the first information-based lower bound for convex feasibility expressed in the Friedrichs angle. Neither is proved in the present paper.

1.4 What is and is not new

Table 1 places the result. We claim the first *global* linear rate with an *explicit* complexity bound for the decentralized *optimization* problem with agent-private constraint sets, in the projection/constrained-solve oracle model. We do not claim the first linear convergence for agent-private sets — Pang [24] obtains an asymptotic linear rate for the feasibility problem, under a local nondegeneracy constraint qualification that neither implies nor is implied by linear regularity — nor the identity (1.2), which is due to Reich and Zalas [27].

The closest complexity results are Yarmoshik et al. [36] and Yarmoshik et al. [35]. The first proves matching bounds $\Theta(\sqrt{\kappa\kappa_A\chi})$ for separable objectives under *affine coupled* constraints $\sum_i(A_i x_i - b_i) = 0$. The second is closer still: its "shared variable" case is $\min_x \sum_i f_i(x)$ subject to $C_i x = c_i$ for every i, which is exactly our problem with $X_i = \{x : C_i x = c_i\}$, and it gives $O(\sqrt{\kappa\hat{\kappa}_{C^\top}\chi}\log\frac{1}{\varepsilon})$ communication rounds with a matching lower bound, where

$$\hat{\kappa}_{C^\top} \;:=\; \frac{\max_i \lambda_{\max}(C_i^\top C_i)}{\lambda_{\min}^{+}\big(\frac{1}{n}\sum_i C_i^\top C_i\big)}. \tag{1.3}$$

The shape is the same as ours. The difference is the oracle, and it is not cosmetic. Their algorithm class may form $C_i^\top C_i x$ but may not solve a linear system in C_i; ours projects onto X_i, which inverts each block exactly. The consequence is a difference in the *quantity* that appears in the bound: $\hat{\kappa}_{C^\top}$ is

	objective	private X_i	gossip	rate	assumption
Nedic et al. [22]	yes	yes	yes	asymptotic	interior point
Lee and Nedic [18]	yes	yes	yes	asymptotic	reg. of own $\bigcap_j X_i^j$
Mai and Abed [20]	yes	yes	yes	$O(\log t/\sqrt{t})$	lin. reg. (in constants)
Shahriari-Mehr and Panahi [31]	yes	yes	yes	$O(1/\sqrt{K})$	none beyond feasibility
Pang [24]	no	yes	yes	asympt. linear	local nondegeneracy CQ
Necoara et al. [21]	no	n/a	no	linear	lin. reg., unaccelerated
Alghunaim et al. [1]	yes	common R	yes	linear	—
Scaman et al. [29]	yes	none	yes	$\sqrt{\kappa\chi}$, optimal	—
Yarmoshik et al. [36]	yes	affine coupled	yes	$\sqrt{\kappa\kappa_A\chi}$, optimal	affine, A_i-oracle
Yarmoshik et al. [35]	yes	affine	yes	$\sqrt{\kappa\hat{\kappa}_{C^\top}\chi}$, optimal	affine, C_i-oracle
this paper	yes	yes (affine)	yes	$\sqrt{\kappa\chi/\sigma}$	lin. reg., projection oracle

Table 1 Position relative to prior work.

attached to the matrices and σ to the sets. Proposition 5.7 makes the relation exact — $1/\sigma$ is the minimum of $\hat{\kappa}_{C^\top}$ over all matrix representations of the same collection $\{X_i\}$, attained when every C_i has orthonormal rows, and the two can differ by an arbitrarily large factor otherwise. Neither bound implies the other: their lower bound does not apply to a projection oracle, and our results say nothing about the matrix-multiplication count, which for us is not a resource. We regard the multiplicative shape of the two upper bounds as mutually corroborating, and say so where it bears on Open Open Problem, stated in the companion paper [12]. Finally, σ is defined for arbitrary closed convex X_i, where no matrix representation and no orthonormalization exist; both of those papers are affine-equality only, and neither mentions linear regularity or Friedrichs angles.

Our Theorem 6.2 does not contradict Alghunaim et al. [1]: linear regularity is an extra assumption their construction is not required to satisfy, and their hard instances live in dimension growing like $\varepsilon^{-1/2}$. We do not claim their family has $\sigma \to 0$, since their r_i need not be indicator functions of a linearly regular collection; what we claim is that adding linear regularity suffices to escape their conclusion, and the lower bound of [12] quantifies the price one pays as σ degrades.

1.5 Notation

$\mathbb{R}^d$ carries the Euclidean inner product and norm. For nonempty closed convex C, P_C is the projection, $\operatorname{dist}(x, C) = \|x - P_C x\|$, $N_C(x)$ the normal cone, ι_C the indicator. For symmetric positive semidefinite M, $\lambda_{\max}(M)$ is the largest and $\lambda^+_{\min}(M)$ the smallest *nonzero* eigenvalue, $\operatorname{ran} M$ the range. T_k is the Chebyshev polynomial of the first kind. For closed subspaces U, V of a finite-dimensional space with $M = U \cap V$,

$$c_F(U, V) = \max\left\{|\langle u, v\rangle| : u \in U \cap M^\perp,\ v \in V \cap M^\perp,\ \|u\| \le 1,\ \|v\| \le 1\right\}.$$

2 Setting

Assumption 2.1 (Objective) Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is differentiable, μ-strongly convex and L-smooth with $0 < \mu \leq L < \infty$; $\kappa := L/\mu \geq 1$.

Assumption 2.2 (Constraints) Each $X_i \subseteq \mathbb{R}^d$ is nonempty, closed and convex, and $X = \bigcap_{i=1}^n X_i \neq \emptyset$.

We split the network hypothesis, because the two halves are used in different places and the Chebyshev operator of Section 6 satisfies only the first.

Assumption 2.3 (Network: spectral) $W \in \mathcal{S}_+^n$ satisfies $\ker W = \operatorname{span}\{\mathbf{1}_n\}$ and $\lambda_{\max}(W) = 1$. We set $\gamma := \lambda_{\min}^+(W) \in (0, 1]$ and $\chi := 1/\gamma$.

Assumption 2.4 (Network: sparsity) $W_{ij} = 0$ whenever $i \neq j$ and $\{i, j\} \notin E$.

A *communication round* is one application of W. Assumptions 2.3–2.4 together are the standard gossip model; γ is the eigengap of Scaman et al. [29], ranging from $\Theta(1)$ for expanders to $\Theta(n^{-2})$ for a path. Note that W is a *disagreement* operator, not an averaging one: $\ker W = \operatorname{span}\{\mathbf{1}\}$ and $\langle x, Wx\rangle = 0$ precisely at consensus, so W plays the role of a normalized graph Laplacian, and the associated mixing matrix is $I - W$. The extreme case $\gamma = 1$ means $W = P_{\mathbf{1}^\perp}$, hence $I - W$ is the consensus projector and a single round achieves exact consensus; it is admissible and handled explicitly wherever it matters.

Product space. Let $\mathcal{H} := (\mathbb{R}^d)^n$ with the *averaged* inner product $\langle x, y\rangle := \frac{1}{n}\sum_i \langle x_i, y_i\rangle$, which differs from the Euclidean inner product on $\mathbb{R}^{nd}$ by a positive scalar, so orthogonal projections, orthogonal complements and principal angles are unchanged. Put

$$F(x) := \frac{1}{n}\sum_{i=1}^n f_i(x_i), \quad \mathcal{X} := X_1 \times \cdots \times X_n, \quad \mathcal{C} := \{(z, \ldots, z) : z \in \mathbb{R}^d\}.$$

Then $\nabla F(x) = (\nabla f_i(x_i))_i$ and F is μ-strongly convex and L-smooth on $\mathcal{H}$ with the same constants as the f_i. The map $\mathrm{J} : \mathbb{R}^d \to \mathcal{C}$, $\mathrm{J}(z) = (z, \ldots, z)$, is an *isometry* onto $\mathcal{C}$, and $P_{\mathcal{C}} x = \mathrm{J}(\bar{x})$ with $\bar{x} = \frac{1}{n}\sum_i x_i$. Problem (P) is equivalent to

$$\min_{x \in \mathcal{H}} F(x) \quad \text{subject to} \quad x \in \mathcal{X} \cap \mathcal{C}. \tag{$\mathcal{P}$}$$

Let $\mathbf{W} : \mathcal{H} \to \mathcal{H}$, $(\mathbf{W}x)_i = \sum_j W_{ij} x_j$; it is self-adjoint and positive semidefinite on $\mathcal{H}$, $\ker \mathbf{W} = \mathcal{C}$, $\lambda_{\max}(\mathbf{W}) = 1$, $\lambda_{\min}^+(\mathbf{W}) = \gamma$.

3 The constraint gap

This section fixes the object the paper is about and records its spectral reading. Propositions 3.4, 3.7 and 3.9(b) are, for affine sets, restatements of results of Reich and Zalas [27]; we say so precisely at each point.

Definition 3.1 (Constraint gap and averaged modulus) For a collection satisfying Assumption 2.2 put

$$\sigma := \inf_{x\in\mathbb{R}^d\setminus X} \frac{\frac{1}{n}\sum_{i=1}^n \operatorname{dist}^2(x, X_i)}{\operatorname{dist}^2(x, X)}, \qquad \theta_2 := \inf\Big\{\theta : \operatorname{dist}(x, X) \le \theta\Big(\tfrac{1}{n}\sum_i \operatorname{dist}^2(x, X_i)\Big)^{1/2}\ \forall x\Big\}.$$

For a bounded B meeting $\mathbb{R}^d \setminus X$, σ_B is the same infimum over $x \in B \setminus X$.

Remark 3.2 (The trivial case $X = \mathbb{R}^d$) If $X = \mathbb{R}^d$ — equivalently $X_i = \mathbb{R}^d$ for every i, the unconstrained problem — both infima in Definition 3.1 are over an empty set, and we adopt the conventions

$$\sigma := 1, \qquad \theta_2 := 1, \qquad \lambda_{\max}\big(\,\cdot\,|_{\{0\}}\big) := 0, \qquad \lambda^+_{\min}\big(\,\cdot\,|_{\{0\}}\big) := 1,$$

under which every identity below evaluates correctly. *Every statement in this section that involves θ_2^{-2}, a restriction to $(X^0)^\perp$, or the characterization (3.2) is to be read with $X \neq \mathbb{R}^d$ unless these conventions are invoked*; we flag the two places where the reading matters. This is a genuinely degenerate case for the geometry — $(X^0)^\perp = \{0\}$, so there is nothing to measure — but it is not degenerate for the paper, since it is the unconstrained sanity check $\sigma = 1$.

Assume $X \neq \mathbb{R}^d$ for the remainder of this subsection. By construction $\sigma = \theta_2^{-2}$ *exactly*: θ_2 is the averaged ℓ_2 linear regularity modulus, and σ is its reciprocal square. Since $X \subseteq X_i$ gives $\operatorname{dist}(x, X_i) \le \operatorname{dist}(x, X)$, we have $\sigma \in [0, 1]$, with $\sigma = 1$ if and only if $X_i = X$ for every i.

Proposition 3.3 (Relation to the conventional modulus) *$\sigma > 0$ if and only if $\{X_1, \dots, X_n\}$ is linearly regular in the sense of Bauschke and Borwein [3], i.e. $\operatorname{dist}(x, X) \le \theta \max_i \operatorname{dist}(x, X_i)$ for all x and some $\theta < \infty$. The optimal such θ satisfies $\theta \le \theta_2 \le \sqrt{n}\,\theta$, equivalently $(n\theta^2)^{-1} \le \sigma \le \theta^{-2}$.*

Proof With $d_i = \operatorname{dist}(x, X_i)$, $d = \operatorname{dist}(x, X)$: from $\frac{1}{n}\sum_i d_i^2 \le \max_i d_i^2$ we get $d \le \theta_2 \max_i d_i$, so $\theta \le \theta_2$; from $\max_i d_i^2 \le n \cdot \frac{1}{n}\sum_i d_i^2$ we get $d \le \theta\sqrt{n}(\frac{1}{n}\sum_i d_i^2)^{1/2}$, so $\theta_2 \le \sqrt{n}\theta$. □

3.1 σ as a spectral gap (affine case)

Proposition 3.4 (Identification; 27, Thm. 8, Rem. 9) *Let $X_1, \dots, X_n$ be closed affine sets with $X \neq \emptyset$ and $X \neq \mathbb{R}^d$ (see Remark 3.2), let X_i^0, X^0 be the direction subspaces, let $\bar{P} := \frac{1}{n}\sum_i P_{X_i^0}$, and let $\mathcal{X}_0 = X_1^0 \times \cdots \times X_n^0$*

and $\mathcal{C}$ be as in Section 2. Then $\bar{P}$ leaves X^0 and $(X^0)^\perp$ invariant, acts as the identity on X^0, and

$$\sigma \;=\; 1 - \lambda_{\max}\big(\bar{P}|_{(X^0)^\perp}\big) \;=\; 1 - \big\|\bar{P} - P_{X^0}\big\| \;=\; 1 - c_F(\mathcal{X}_0, \mathcal{C})^2.$$

If moreover $n \geq 2$ and $X_i^0 \neq X^0$ for at least one i — equivalently $\sigma < 1$, the only case in which the supremum below is over a nonempty set — then also

$$\sigma \;=\; \tfrac{n-1}{n}\big(1 - c_{\mathrm{RZ}}(X_1^0, \dots, X_n^0)\big),$$

where c_{RZ} is the Friedrichs number of the collection in the sense of Reich and Zalas [27, Def. 7].

Remark 3.5 (Two different "Friedrichs cosines") The last equality is worth isolating because the literature uses one name for two quantities. Throughout this paper $c_F(\mathcal{X}_0, \mathcal{C})$ is the Friedrichs cosine of the *pair* $(\mathcal{X}_0, \mathcal{C})$ in Pierra's product space, whereas $c_{\mathrm{RZ}}(X_1^0, \dots, X_n^0)$ is the Friedrichs number of the *collection*, defined in Reich and Zalas [27, Def. 7] as

$$c_{\mathrm{RZ}} := \sup\Big\{ \tfrac{1}{n-1} \tfrac{\sum_{i \neq j} \langle z_i, z_j \rangle}{\sum_i \|z_i\|^2} : z_i \in X_i^0 \cap (X^0)^\perp,\ \textstyle\sum_i \|z_i\|^2 \neq 0 \Big\}.$$

(The constraint set is $\{0\}$ exactly when $X_i^0 = X^0$ for every i, i.e. when $\sigma = 1$; that degenerate case is excluded in Proposition 3.4, where the first three expressions remain valid.) They are related by $c_F(\mathcal{X}_0, \mathcal{C})^2 = \frac{n-1}{n} c_{\mathrm{RZ}} + \frac{1}{n}$ [27, Rem. 9], which is where the displayed fourth expression comes from. In particular $\sigma = 1 - c_F^2$, whereas $1 - c_{\mathrm{RZ}}$ exceeds σ by the factor $\frac{n}{n-1}$; for $n = 2$ that factor is 2, and c_{RZ} reduces to the classical two-subspace Friedrichs cosine $\cos\theta_F$, giving $\sigma = \frac{1}{2}(1 - \cos\theta_F)$ as in Proposition 3.9(b). Substituting one cosine for the other is an easy error to make and changes every constant in this paper by $\frac{n}{n-1}$.

Proof Translating by a point of X we may take every X_i to be a subspace. For $z \in X$ we have $P_{X_i} z = z$ for all i, so $\bar{P} z = z$; by self-adjointness $X^\perp$ is invariant. For any z,

$$\frac{1}{n} \sum_{i=1}^{n} \mathrm{dist}^2(z, X_i) = \frac{1}{n} \sum_i \langle z, (I - P_{X_i}) z \rangle = \langle z, (I - \bar{P}) z \rangle. \tag{3.1}$$

Decomposing $z = z_X + z_\perp$ with $z_X \in X \subseteq X_i$ gives $\mathrm{dist}(z, X_i) = \mathrm{dist}(z_\perp, X_i)$ and $\mathrm{dist}(z, X) = \|z_\perp\|$, so the quotient in Definition 3.1 depends on z only through $z_\perp$ and

$$\sigma = \min_{0 \neq z \in X^\perp} \frac{\langle z, (I - \bar{P}) z \rangle}{\|z\|^2} = 1 - \lambda_{\max}(\bar{P}|_{X^\perp}).$$

Since $\bar{P} - P_X$ is self-adjoint, positive semidefinite and vanishes on X, $\lambda_{\max}(\bar{P}|_{X^\perp}) = \|\bar{P} - P_X\|$. The third equality is Reich and Zalas [27, Thm. 8] at $k = 1$ (see also 28, Thms. 1.6, 3.3, and 27, Cor. 15 for the affine case), which states $\|\bar{P} - P_X\| = c_F(\mathcal{X}_0, \mathcal{C})^2$; a self-contained proof is obtained by applying the characterization, valid for closed subspaces U, V with $V \not\subseteq U$,

$$1 - c_F(U,V)^2 = \min\left\{ \tfrac{\operatorname{dist}^2(v,U)}{\|v\|^2} : v \in V \ominus (U \cap V),\ v \neq 0 \right\} \tag{3.2}$$

with $U = \mathcal{X}_0$, $V = \mathcal{C}$ — the hypothesis $\mathcal{C} \not\subseteq \mathcal{X}_0$ is exactly $X \neq \mathbb{R}^d$ (Remark 3.2) — using that J is an isometry carrying X onto $\mathcal{X}_0 \cap \mathcal{C}$ and that $\operatorname{dist}^2(\mathrm{J}(z), \mathcal{X}_0) = \frac{1}{n}\sum_i \operatorname{dist}^2(z, X_i)$. The fourth equality follows from the third and $c_F(\mathcal{X}_0, \mathcal{C})^2 = \frac{n-1}{n} c_{\mathrm{RZ}} + \frac{1}{n}$ [27, Rem. 9]. □

Remark 3.6 (What is quoted and what is not) Reich and Zalas and Reich and Zalas [27, 28] prove the *operator-norm* identity $\|\bar{P}^k - P_{X^0}\| = c_F(\mathcal{X}_0, \mathcal{C})^{2k}$ (and its arbitrary-weight, countable-family version). The *distance* form $\sigma = 1 - \|\bar{P} - P_{X^0}\|$ — that is, the statement that the best constant in the averaged error bound equals the spectral gap — is the one-line consequence (3.1) above, which we spell out only because we use it constantly. We do not claim either as a contribution of this paper.

The next statement is the one that makes "spectral gap" operational, and it is where affine sets must be handled with care: $\frac{1}{n}\sum_i P_{X_i}$ is an *affine* map, not a linear one, and it is that affine map — not $\bar{P}$ — that contracts towards X.

Proposition 3.7 (Simultaneous projection) *In the setting of Proposition 3.4 let $Q(z) := \frac{1}{n}\sum_{i=1}^n P_{X_i}(z)$. Fixing any $x^0 \in X$ one has $Q(z) = x^0 + \bar{P}(z - x^0)$, and for every z and every $k \in \mathbb{N}$*

$$\operatorname{dist}\left(Q^k(z), X\right) \ \le\ (1-\sigma)^k \operatorname{dist}(z, X).$$

The factor $1 - \sigma$ is the exact asymptotic rate whenever $z - x^0$ has a nonzero component along the top eigenspace of $\bar{P}|_{(X^0)^\perp}$.

Proof For $x^0 \in X \subseteq X_i$, $P_{X_i}(z) = x^0 + P_{X_i^0}(z - x^0)$; averaging gives $Q(z) = x^0 + \bar{P}(z - x^0)$, so $Q^k(z) = x^0 + \bar{P}^k(z - x^0)$. With $u = z - x^0$,

$$\begin{aligned}\operatorname{dist}(Q^k(z), X) &= \operatorname{dist}(\bar{P}^k u, X^0) = \|P_{(X^0)^\perp}\bar{P}^k u\| = \|\bar{P}^k P_{(X^0)^\perp} u\| \\ &\le (1-\sigma)^k \|P_{(X^0)^\perp} u\| = (1-\sigma)^k \operatorname{dist}(z, X),\end{aligned}$$

using invariance of $(X^0)^\perp$ and $\|\bar{P}|_{(X^0)^\perp}\| = 1 - \sigma$ from Proposition 3.4. Exactness of the rate is the standard power iteration statement. □

Remark 3.8 Taking z arbitrary and $X_1 = \cdots = X_n = \{p\}$ shows why the affine form is needed: there $\bar{P} = 0$ and $\sigma = 1$, and $Q(z) = p$ for every z, whereas the linear iteration $z \mapsto \bar{P}z$ sends z to $0 \notin X$.

	network gap γ	**constraint gap** σ
residual operator	$\mathbf{W}$ (disagreement)	$I - \bar{P}$ (infeasibility)
averaging operator	$I - \mathbf{W}$ (one gossip round)	$\bar{P} = Q$ (one simultaneous projection)
kernel of the residual	consensus subspace $\mathcal{C}$	feasible set $X = X^0$
gap	$\gamma = \lambda^+_{\min}(\mathbf{W})$	$\sigma = \lambda^+_{\min}\big((I-\bar{P})\|_{X^\perp}\big) = 1 - \lambda_{\max}(\bar{P}\|_{X^\perp})$
one averaging step	contracts by $1-\gamma$ towards $\mathcal{C}$	contracts by $1-\sigma$ towards X
error bound	$\operatorname{dist}^2(x,\mathcal{C}) \le \gamma^{-1}\langle x, \mathbf{W}x\rangle$	$\operatorname{dist}^2(x,X) \le \sigma^{-1}\frac{1}{n}\sum_i \operatorname{dist}^2(x,X_i)$
residual energy	$\langle x, \mathbf{W}x\rangle$	$\frac{1}{n}\sum_i \operatorname{dist}^2(x,X_i) = \langle x, (I-\bar{P})x\rangle$
degenerate case	disconnected graph	$\{X_i\}$ not linearly regular

Table 2 The two gaps. In both columns the *residual* operator is the one whose kernel is the target set and whose smallest nonzero eigenvalue is the gap; the corresponding *averaging* operator is I minus it, and contracts towards the target at rate one minus the gap. Thus $\mathbf{W}$ plays the role of a graph Laplacian (it measures disagreement, it does not average), and its counterpart on the constraint side is $I - \bar{P}$, not $\bar{P}$. *The right column is stated for linear* X_i, i.e. after translating an arbitrary affine collection so that $0 \in X$; then $X = X^0$ and the simultaneous projection Q coincides with the linear map $\bar{P}$. Without that normalization Q is affine and $Q(z) = x^0 + \bar{P}(z - x^0)$ for any $x^0 \in X$ (Proposition 3.7); it is Q, not $\bar{P}$, that contracts towards X.

3.2 Computing and estimating σ

Proposition 3.9

(a) (Hyperplanes.) *If* $X_i = \{x : \langle a_i, x\rangle = b_i\}$ *with* $\|a_i\| = 1$ *and* $X \neq \emptyset$, *then* $\sigma = \lambda^+_{\min}\big(\frac{1}{n}\sum_i a_i a_i^\top\big)$.

(b) (Two sets.) *If* $n = 2$, X_1, X_2 *are affine with* $X_1 \cap X_2 \neq \emptyset$ *and* $X_1^0 \neq X_2^0$, *and* θ_F *is the Friedrichs angle between the direction subspaces, then* $\sigma = \frac{1}{2}(1 - \cos\theta_F) = \sin^2(\theta_F/2)$, *so* $\sigma^{-1/2} = \Theta(1/\sin\theta_F)$ *as* $\theta_F \downarrow 0$. (This is 27, Rem. 11.)

(c) (Polyhedra.) *If every* X_i *is polyhedral and* $X \neq \emptyset$ *then* $\sigma > 0$, *and the regularity is global rather than merely bounded.*

(d) (Interior condition.) *If* $X_n \cap \operatorname{int}(X_1 \cap \cdots \cap X_{n-1}) \neq \emptyset$ *then* $\sigma_B > 0$ *for every bounded* B.

Proof (a) Fix $x^0 \in X$. Then $\operatorname{dist}(x, X_i) = |\langle a_i, x - x^0\rangle|$, so with $w = x - x^0$ and $G = \frac{1}{n}\sum_i a_i a_i^\top$ we get $\frac{1}{n}\sum_i \operatorname{dist}^2(x,X_i) = \langle w, G\rangle w$, while $X = x^0 + (\operatorname{ran} G)^\perp$ gives $\operatorname{dist}^2(x,X) = \|P_{\operatorname{ran} G} w\|^2$; take the infimum over $w \in \operatorname{ran} G \setminus \{0\}$.

(b) We may take X_1, X_2 to be subspaces. By the two-subspace (Halmos) decomposition, $(X_1 \cap X_2)^\perp$ splits orthogonally into two-dimensional blocks, on which $P_{X_1} + P_{X_2}$ has eigenvalues $1 \pm \cos t_j$ with t_j the principal angles, together with the three exceptional summands $X_1 \cap X_2^\perp$ and $X_1^\perp \cap X_2$, on which $P_{X_1} + P_{X_2}$ acts as the identity, and $X_1^\perp \cap X_2^\perp$, on which it vanishes. Hence $\lambda_{\max}(\bar{P}|_{X^\perp}) \in \{\frac{1+\cos\theta_F}{2}, \frac{1}{2}, 0\}$ and equals the largest of those present. Since $X_1 \neq X_2$, one of $X_1 \ominus X, X_2 \ominus X$ is nonzero; a unit u in it satisfies $\langle u, \bar{P}u\rangle = \frac{1}{2}(1 + \|P_{X_2}u\|^2) \geq \frac{1}{2}$, so the value 0 is not the maximum and $\lambda_{\max}(\bar{P}|_{X^\perp}) = \frac{1+c_F}{2}$, where $c_F = 0$ precisely when no generic block is present. Proposition 3.4 gives the claim; the asymptotics follow from $\sin^2\theta_F = (1 - \cos\theta_F)(1 + \cos\theta_F)$.

(c) Hoffman's error bound [11] applied to the linear system defining X yields a single constant valid on all of $\mathbb{R}^d$; combine with Proposition 3.3. (d) is Bauschke and Borwein [3, Cor. 5.14]. $\square$

Example 3.10 (Nearly redundant local constraints) Let $\|a_0\| = 1$ and let $u_1, \ldots, u_n$ be unit vectors that span $a_0^\perp$ and satisfy $\sum_i u_i = 0$ (take them in $\pm$ pairs, so $n \ge 2(d-1)$). Put $a_i = (a_0 + \varepsilon u_i)/\|a_0 + \varepsilon u_i\|$ and $X_i = \{x : \langle a_i, x\rangle = \langle a_i, x^0\rangle\}$, so that $x^0 \in X$ and $X \neq \emptyset$. Since $u_i \perp a_0$ and $\|u_i\| = 1$ we have $\|a_0 + \varepsilon u_i\| = \sqrt{1+\varepsilon^2}$ for every i, so with $G_\varepsilon := \frac{1}{n}\sum_i a_i a_i^\top$ and $\Sigma := \frac{1}{n}\sum_i u_i u_i^\top$,

$$G_\varepsilon = \frac{1}{1+\varepsilon^2}\Big(a_0 a_0^\top + \frac{\varepsilon}{n}\big(a_0 {\textstyle\sum_i} u_i^\top + {\textstyle\sum_i} u_i a_0^\top\big) + \varepsilon^2 \Sigma\Big) = \frac{1}{1+\varepsilon^2}\big(a_0 a_0^\top + \varepsilon^2 \Sigma\big),$$

the cross terms vanishing because $\sum_i u_i = 0$. This is block diagonal for $\mathbb{R}a_0 \oplus a_0^\perp$, and Σ is positive definite on $a_0^\perp$ because the u_i span it, so for ε small

$$\sigma \;=\; \lambda^+_{\min}(G_\varepsilon) \;=\; \frac{\varepsilon^2}{1+\varepsilon^2}\,\lambda_{\min}\big(\Sigma|_{a_0^\perp}\big) \;=\; \Theta(\varepsilon^2) \qquad (\varepsilon \downarrow 0)$$

by Proposition 3.9(a). Every individual constraint stays perfectly conditioned while the collection degenerates: σ is invisible to any hypothesis stated one set at a time. The centring $\sum_i u_i = 0$ is not cosmetic — with $u_1 = \cdots = u_n$ every X_i coincides and $\sigma = 1$ for every ε — but it is the generic situation: a first-order cross term shifts $\lambda^+_{\min}(G_\varepsilon)$ by $O(\varepsilon^2)$ as well, provided the $\langle u_i, v\rangle$ are not all equal.

4 Optimality conditions and stability

Two consequences of $\sigma > 0$, valid for general closed convex sets, which explain why the problem is decentralizable at all.

Theorem 4.1 (Strong CHIP and separated optimality) *Let Assumptions 2.1–2.2 hold with $\sigma > 0$, and let $\bar{x} \in X$. Then*

$$N_X(\bar{x}) = \sum_{i=1}^n N_{X_i}(\bar{x}), \tag{4.1}$$

and $\bar{x}$ is the (unique) solution of (P) if and only if there exist $v_i \in N_{X_i}(\bar{x})$, $i = 1, \ldots, n$, with

$$\frac{1}{n}\sum_{i=1}^n \big(\nabla f_i(\bar{x}) + v_i\big) = 0. \tag{4.2}$$

Proof $\sigma > 0$ is linear regularity, hence bounded linear regularity, which implies the strong CHIP [4, Thm. 3]; this is (4.1) ($\supseteq$ always holds). Given (4.1), first-order optimality $0 \in \nabla f(\bar{x}) + N_X(\bar{x})$ for the μ-strongly convex f over the convex set X is necessary and sufficient for $\bar{x}$ to solve (P), and is equivalent to the existence of $w_i \in N_{X_i}(\bar{x})$ with $\nabla f(\bar{x}) + \sum_i w_i = 0$; setting $v_i = n w_i$, which again lies in the cone $N_{X_i}(\bar{x})$, gives (4.2). Uniqueness follows from strong convexity. $\square$

Condition (4.2) is a certificate no agent could verify alone but that every agent can contribute to: agent i supplies $(\nabla f_i(\bar{x}), v_i)$ from its own data, and the agents need only agree that the average vanishes. Without the strong CHIP, N_X can be strictly larger than $\sum_i N_{X_i}$ and no such separation exists.

Theorem 4.2 (Perturbation bound) *Let $\tilde{f}_i$ satisfy Assumption 2.1 with the same μ, L, let $\tilde{X}_i$ satisfy Assumption 2.2, write $\tilde{f} = \frac{1}{n}\sum_i \tilde{f}_i$, $\tilde{X} = \bigcap_i \tilde{X}_i$, and let $x^\star, \tilde{x}^\star$ be the two solutions. Let B be a bounded convex set containing $x^\star, \tilde{x}^\star$ and the segments joining them to their projections onto X and $\tilde{X}$, and suppose*

$$\sup_{x\in B}\|\nabla f(x) - \nabla \tilde{f}(x)\| \le \delta_f, \qquad \max_i \sup_{x\in B}\left|\operatorname{dist}(x, X_i) - \operatorname{dist}(x, \tilde{X}_i)\right| \le \delta_X,$$

with both collections having constraint gap at least $\sigma > 0$ on B. Put $\Gamma := \sup_{x\in B}\max\{\|\nabla f(x)\|, \|\nabla \tilde{f}(x)\|\}$, which is finite. Then

$$\|\tilde{x}^\star - x^\star\| \;\le\; \frac{2\delta_f}{\mu} \;+\; \sqrt{\frac{2(2\Gamma + \delta_f)\,\delta_X}{\mu\sqrt{\sigma}}} \;+\; \frac{\delta_X}{\sqrt{\sigma}}.$$

Proof Since $\tilde{x}^\star \in \tilde{X}_i$ we get $\operatorname{dist}(\tilde{x}^\star, X_i) \le \delta_X$ for all i, so Definition 3.1 gives $\operatorname{dist}(\tilde{x}^\star, X) \le \delta_X/\sqrt{\sigma}$; symmetrically $\operatorname{dist}(x^\star, \tilde{X}) \le \delta_X/\sqrt{\sigma}$. Put $y = P_X(\tilde{x}^\star)$ and $\tilde{y} = P_{\tilde{X}}(x^\star)$. Strong convexity and optimality of $x^\star$ over $X \ni y$ give $f(y) \ge f(x^\star) + \frac{\mu}{2}\|y - x^\star\|^2$. Also $f(y) \le f(\tilde{x}^\star) + \Gamma\delta_X/\sqrt{\sigma}$, and with $h := f - \tilde{f}$ (so $|h(u) - h(v)| \le \delta_f\|u - v\|$ on B), optimality of $\tilde{x}^\star$ over $\tilde{X} \ni \tilde{y}$ gives

$$f(\tilde{x}^\star) = \tilde{f}(\tilde{x}^\star) + h(\tilde{x}^\star) \le \tilde{f}(\tilde{y}) + h(\tilde{x}^\star) \le f(x^\star) + \Gamma\tfrac{\delta_X}{\sqrt{\sigma}} + h(\tilde{x}^\star) - h(x^\star).$$

Combining, $\frac{\mu}{2}\|y - x^\star\|^2 \le \delta_f\|\tilde{x}^\star - x^\star\| + 2\Gamma\delta_X/\sqrt{\sigma}$. Writing $a = \|\tilde{x}^\star - x^\star\|$, $b = \|y - x^\star\|$ and using $a \le b + \delta_X/\sqrt{\sigma}$ yields $\frac{\mu}{2}b^2 - \delta_f b \le (2\Gamma + \delta_f)\delta_X/\sqrt{\sigma}$; solving the quadratic and using $\sqrt{p+q} \le \sqrt{p} + \sqrt{q}$ gives $b \le \frac{2\delta_f}{\mu} + \sqrt{2(2\Gamma + \delta_f)\delta_X/(\mu\sqrt{\sigma})}$, and $a \le b + \delta_X/\sqrt{\sigma}$. □

The bound is Lipschitz in the objective perturbation with modulus $2/\mu$, and exhibits a $\sigma^{-1/4}$ amplification in the Hölder term together with a $\sigma^{-1/2}$ amplification in the feasibility term. We do not claim these exponents are unimprovable; both are inherited from the error bound of Definition 3.1, and establishing matching lower examples is left open.

5 Dual conditioning

Assumption 5.1 (Affine local sets) Each X_i is a nonempty closed affine subset of $\mathbb{R}^d$ with direction subspace X_i^0. We fix $x^0 \in X$ and write $\mathcal{X} = \mathrm{J}(x^0) + \mathcal{X}_0$ with $\mathcal{X}_0 = X_1^0 \times \cdots \times X_n^0$.

Assumption 5.1 holds for the rest of this section and all of Section 6; Section 6.5 discusses what is known and open without it. Under Assumption 5.1 one has $\sigma > 0$ automatically by Proposition 3.9(c), so the content of what follows is entirely quantitative. Affine local constraints already cover linear measurement and consistency relations, conservation laws, and subspace membership.

We dualize *consensus only*, leaving each local set in the primal so that no agent ever needs another agent's constraint. Let $\mathbf{A} := \mathbf{W}^{1/2}$, so $\ker \mathbf{A} = \mathcal{C}$ and $x \in \mathcal{C} \iff \mathbf{A}x = 0$, and set

$$G := F + \iota_{\mathcal{X}}, \qquad \Psi(y) := G^*(-\mathbf{A}y), \qquad y \in \mathcal{C}^\perp.$$

Minimizing Ψ over $\mathcal{C}^\perp$ is the dual of ($\mathcal{P}$); strong duality holds because the coupling constraint is a linear equation with a feasible point and $\operatorname{ri}\mathcal{X} = \mathcal{X}$ under Assumption 5.1.

5.1 A spectral lemma

Lemma 5.2 *Let $\mathcal{H}_0$ be a finite-dimensional Hilbert space, A self-adjoint positive definite on $\mathcal{H}_0$, and R self-adjoint positive semidefinite, $R \neq 0$. Then*

$$\ker(ARA) = A^{-1}(\ker R), \qquad \operatorname{ran}(ARA) = A\operatorname{ran}(R), \qquad \lambda^+_{\min}(ARA) \geq \lambda_{\min}(A)^2\,\lambda^+_{\min}(R).$$

Proof Let P be the orthogonal projector onto $\operatorname{ran} R$. Since A is injective, $ARAu = 0 \iff RAu = 0 \iff Au \in \ker R = \ker P \iff APAu = 0$; thus ARA and APA have the same kernel and, being self-adjoint, the same range $V := A\operatorname{ran}(R)$. The nonzero eigenvalues of $APA = (AP)(PA)$ coincide with those of $(PA)(AP) = PA^2P$, and $\lambda^+_{\min}(PA^2P) = \min\{\langle u, A^2u\rangle/\|u\|^2 : 0 \neq u \in \operatorname{ran} P\} \geq \lambda_{\min}(A)^2$. Finally $R \succeq \lambda^+_{\min}(R)P$ gives $ARA \succeq \lambda^+_{\min}(R)APA$, and since V is invariant for both, $\lambda^+_{\min}(ARA) \geq \lambda^+_{\min}(R)\lambda^+_{\min}(APA) \geq \lambda^+_{\min}(R)\lambda_{\min}(A)^2$. □

5.2 The main estimate

Theorem 5.3 (Dual conditioning) *Let Assumptions 2.1, 2.2, 2.3 and 5.1 hold, and suppose $\mathcal{X}_0 \not\subseteq \mathcal{C}$. Let E be a linear isometry onto $\mathcal{X}_0$ with $E^*E = I$ and $EE^* = P_{\mathcal{X}_0}$* (with respect to the averaged inner product)*, and set $B := -E^*\mathbf{A}|_{\mathcal{C}^\perp}$, $\mathcal{Y} := \mathcal{C}^\perp \ominus \ker B$. Then*

(i) *Ψ is convex and $\frac{1}{\mu}$-smooth on $\mathcal{C}^\perp$;*

(ii) *$\nabla\Psi(y) \in \mathcal{Y}$ for every $y \in \mathcal{C}^\perp$, and $\Psi|_{\mathcal{Y}}$ is $\frac{\sigma\gamma}{L}$-strongly convex;*

(iii) *consequently* $\kappa_{\text{dual}} := \dfrac{\textit{smoothness of } \Psi|_{\mathcal{Y}}}{\textit{strong convexity of } \Psi|_{\mathcal{Y}}} \leq \dfrac{\kappa\chi}{\sigma}.$

Moreover $x(y) := \arg\min_{x\in\mathcal{X}}\{F(x) + \langle \mathbf{A}y, x\rangle\}$ depends on y only through $P_{\mathcal{Y}}y$ and is $\frac{1}{\mu}$-Lipschitz.

Proof Reduction. Writing $x = \mathrm{J}(x^0) + Eu$ and $\varphi(u) := F(\mathrm{J}(x^0) + Eu)$, the function φ is μ-strongly convex and L-smooth because E is an isometry, so φ^* is $\frac{1}{L}$-strongly convex and $\frac{1}{\mu}$-smooth. For any $z \in \mathcal{H}$, $G^*(z) = \sup_{x\in\mathcal{X}}\{\langle z, x\rangle - F(x)\} = \langle z, \mathrm{J}(x^0)\rangle + \varphi^*(E^*z)$, so with $c := -\mathbf{A}\mathrm{J}(x^0)$,

$$\Psi(y) = \langle c, y\rangle + \varphi^*(By), \qquad y \in \mathcal{C}^\perp. \tag{5.1}$$

(i). $\|\mathbf{A}\|^2 = \lambda_{\max}(\mathbf{W}) = 1$, so $\|B\| \le 1$ and Ψ is $\frac{1}{\mu}\|B\|^2 \le \frac{1}{\mu}$-smooth.

(ii). $\nabla\Psi(y) = c + B^*\nabla\varphi^*(By)$ and $\operatorname{ran} B^* = (\ker B)^\perp \cap \mathcal{C}^\perp = \mathcal{Y}$. Moreover $c = -\mathbf{A}\mathrm{J}(x^0) = 0$, because $\mathrm{J}(x^0) \in \mathcal{C} = \ker \mathbf{A}$; hence $\nabla\Psi(y) \in \mathcal{Y}$. (More generally, for any coupling for which $c \ne 0$ one notes that $\Psi(y + tu) = \Psi(y) + t\langle c, u\rangle$ for $u \in \ker B \cap \mathcal{C}^\perp$, so boundedness below — guaranteed by strong duality — forces $\langle c, u\rangle = 0$ and again $c \in \mathcal{Y}$.)

For strong convexity it suffices, by $\frac{1}{L}$-strong convexity of φ^*, to show $\|Bw\|^2 \ge \sigma\gamma\|w\|^2$ for $w \in \mathcal{Y}$. Now $\|Bw\|^2 = \langle w, \mathbf{A}P_{\mathcal{X}_0}\mathbf{A}w\rangle$, and since $\mathbf{A}$ maps $\mathcal{C}^\perp$ into itself this operator equals ARA on $\mathcal{C}^\perp$ with $A := \mathbf{A}|_{\mathcal{C}^\perp}$ (positive definite, $\lambda_{\min}(A)^2 = \gamma$) and $R := P_{\mathcal{C}^\perp}P_{\mathcal{X}_0}P_{\mathcal{C}^\perp}|_{\mathcal{C}^\perp}$, which is nonzero because $\mathcal{X}_0 \not\subseteq \mathcal{C}$. Lemma 5.2 gives $\mathcal{Y} = \operatorname{ran}(ARA)$ and $\lambda^+_{\min}(ARA) \ge \gamma\lambda^+_{\min}(R)$.

It remains to identify $\lambda^+_{\min}(R)$. We have $\ker R = \mathcal{C}^\perp \cap \mathcal{X}_0^\perp$ and, for $u \in \mathcal{C}^\perp \ominus \ker R$, $\langle u, R\rangle u = \|P_{\mathcal{X}_0}u\|^2 = \operatorname{dist}^2(u, \mathcal{X}_0^\perp)$, so (3.2) applied to the pair $(\mathcal{X}_0^\perp, \mathcal{C}^\perp)$ gives $\lambda^+_{\min}(R) = 1 - c_F(\mathcal{X}_0^\perp, \mathcal{C}^\perp)^2$. By the classical identity $c_F(U, V) = c_F(U^\perp, V^\perp)$ [6, Thm. 16], see also [7, Lem. 9.5], this equals $1 - c_F(\mathcal{X}_0, \mathcal{C})^2 = \sigma$ by Proposition 3.4. Hence $\lambda^+_{\min}(ARA) \ge \sigma\gamma$.

(iii) follows, and the final claim holds because $x(y) = \mathrm{J}(x^0) + E\nabla\varphi^*(By)$ depends on y only through By, with $\nabla\varphi^*$ being $\frac{1}{\mu}$-Lipschitz and $\|B\| \le 1$. □

Remark 5.4 The excluded case $\mathcal{X}_0 \subseteq \mathcal{C}$ means every $X_i^0 = \{0\}$, i.e. each X_i is a singleton and (P) has no optimization content.

5.3 The bound is attained

The following replaces what was, in an earlier draft of this work, a numerical observation. It shows that no factor in $\kappa\chi/\sigma$ can be removed or improved, and it supplies the hard family used in the lower bounds of [12].

Theorem 5.5 (Extremal family) *Let $\kappa \ge 1$, $\chi \ge 1$, $\sigma \in (0, \frac{1}{2}]$ and let $n \ge 4$ be even. Put $\gamma = 1/\chi$, let $\psi \in (0, \pi/4]$ satisfy $\sin^2\psi = \sigma$, and set $\varepsilon_i = +1$ for $i \le n/2$ and $\varepsilon_i = -1$ otherwise. Let*

$$v := \tfrac{1}{\sqrt{n}}(\varepsilon_1, \dots, \varepsilon_n)^\top,$$

let W be any matrix satisfying Assumption 2.3 with $Wv = \gamma v$ and remaining eigenvalues in $[\gamma, 1]$ (admissible on the complete graph)*, and in $\mathbb{R}^3$ put*

$$a_i := (0,\ \cos\psi,\ \varepsilon_i \sin\psi)^\top, \qquad X_i := \{x \in \mathbb{R}^3 : \langle a_i, x\rangle = 0\}, \qquad f_i(x) := \tfrac{1}{2}\langle x, Hx\rangle - \langle g_i, x\rangle$$

with $H = \operatorname{diag}(\mu, L, L)$*,* $L/\mu = \kappa$*, and arbitrary* g_i*. Then this instance has gossip condition number* χ*, condition number* κ*, constraint gap* σ*, and*

$$\lambda_{\max}(M) = \frac{1}{\mu}, \qquad \lambda^+_{\min}(M) = \frac{\sigma\gamma}{L}, \qquad \kappa_{\text{dual}} = \frac{\kappa\chi}{\sigma},$$

where M *is the operator* $\mathbf{A}E(E^*HE)^{-1}E^*\mathbf{A}$ *of Theorem 5.3 restricted to* $\mathcal{C}^\perp$*.*

Proof Parameters. $G := \frac{1}{n}\sum_i a_i a_i^\top = \operatorname{diag}(0, \cos^2\psi, \sin^2\psi)$ because $\frac{1}{n}\sum_i \varepsilon_i = 0$. Hence $X = \bigcap_i X_i = \operatorname{span}\{e_1\}$ and, by Proposition 3.9(a), $\sigma = \lambda^+_{\min}(G) = \min\{\cos^2\psi, \sin^2\psi\} = \sin^2\psi$ since $\psi \le \pi/4$.

Block structure. $X_i^0 = \operatorname{span}\{e_1, b_i\}$ with $b_i = (0, -\varepsilon_i \sin\psi, \cos\psi)^\top$, and H preserves X_i^0 with $He_1 = \mu e_1$, $Hb_i = Lb_i$. Consequently

$$E(E^*HE)^{-1}E^* \text{ acts blockwise as } \tfrac{1}{\mu}e_1e_1^\top + \tfrac{1}{L}b_ib_i^\top \text{ on agent } i. \tag{5.2}$$

$\lambda_{\max}(M) = 1/\mu$. Let $\mathcal{S}_1 := \{x \in \mathcal{H} : x_i \in \operatorname{span}\{e_1\}\} \cap \mathcal{C}^\perp$. Since $e_1 \in X_i^0$ for all i, (5.2) shows $E(E^*HE)^{-1}E^* = \frac{1}{\mu}I$ on $\mathcal{S}_1$, and $\mathbf{A}$ preserves $\mathcal{S}_1$; hence $M|_{\mathcal{S}_1} = \mathbf{W}|_{\mathcal{S}_1}/\mu$, whose largest eigenvalue is $\lambda_{\max}(W)/\mu = 1/\mu$. With Theorem 5.3(i) this gives equality.

$\lambda^+_{\min}(M) \le \sigma\gamma/L$. Take $u := v \otimes e_2 \in \mathcal{C}^\perp$ (indeed $v \perp \mathbf{1}$). Then $\mathbf{A}u = \sqrt{\gamma}\, v \otimes e_2$, and by (5.2), using $\langle e_1, e_2\rangle = 0$ and $\langle b_i, e_2\rangle = -\varepsilon_i \sin\psi$,

$$\langle u, Mu\rangle = \frac{1}{n}\sum_{i=1}^n \Big\langle \sqrt{\gamma} v_i e_2,\ \tfrac{1}{L}\sqrt{\gamma} v_i(-\varepsilon_i \sin\psi)\, b_i \Big\rangle = \frac{\gamma \sin^2\psi}{L} \cdot \frac{1}{n}\sum_i v_i^2 = \frac{\sigma\gamma}{L}\,\|u\|^2.$$

$u \perp \ker M$. By Lemma 5.2, $\ker M \cap \mathcal{C}^\perp = \mathbf{A}^{-1}(\mathcal{X}_0^\perp \cap \mathcal{C}^\perp)$, and $\mathcal{X}_0^\perp = \prod_i \operatorname{span}\{a_i\}$, so an element of $\mathcal{X}_0^\perp \cap \mathcal{C}^\perp$ has the form $(c_i a_i)_i$ with $\sum_i c_i a_i = 0$, i.e. $\cos\psi \sum_i c_i = 0$ and $\sin\psi \sum_i \varepsilon_i c_i = 0$, i.e. $\sum_{i \le n/2} c_i = \sum_{i > n/2} c_i = 0$. Since $\mathbf{A}$ is self-adjoint and invertible on $\mathcal{C}^\perp$, $u \perp \mathbf{A}^{-1}(c_i a_i)_i$ is equivalent to $\langle \mathbf{A}^{-1}u, (c_i a_i)_i\rangle = 0$, and $\mathbf{A}^{-1}u = \gamma^{-1/2} v \otimes e_2$ gives

$$\langle v \otimes e_2, (c_i a_i)_i\rangle = \frac{\cos\psi}{n}\sum_i v_i c_i = \frac{\cos\psi}{n\sqrt{n}}\Big(\sum_{i \le n/2} c_i - \sum_{i > n/2} c_i\Big) = 0.$$

Hence $u \in \operatorname{ran} M$ and $\lambda^+_{\min}(M) \le \langle u, Mu\rangle/\|u\|^2 = \sigma\gamma/L$; with Theorem 5.3(ii) this is an equality. Dividing gives $\kappa_{\text{dual}} = (1/\mu)/(\sigma\gamma/L) = \kappa\chi/\sigma$. □

Remark 5.6 (Numerical confirmation) Theorem 5.5 was checked on 81 instances spanning $n \in \{4, 6, 10\}$, $\kappa \in \{1, 7, 60\}$, $\chi \in \{1, 4, 25\}$ and three values of ψ; the largest observed value of $|\kappa_{\text{dual}}\sigma/(\kappa\chi) - 1|$ was 1.4×10^{-12} (Section 7). On random instances the bound is far from tight: over 27 instances with $n \in \{6, 10, 14\}$ the ratio $\kappa_{\text{dual}}\sigma/(\kappa\chi)$ lay in $[0.12, 0.30]$, and it falls to 0.018 on the larger instance used in Section 7.

5.4 σ is the representation-free core of the matrix condition number

The affine case of (P) is also treated by Yarmoshik et al. [35] in a matrix-multiplication oracle model, where the constant is $\hat{\kappa}_{C^\top}$ of (1.3). The two constants are related exactly, and the relation explains what the projection oracle buys.

Proposition 5.7 (Minimum over representations) *Let $X_i = \{x \in \mathbb{R}^d : C_i x = c_i\}$ with $X = \bigcap_i X_i \neq \emptyset$ and every $C_i \neq 0$, and let σ be the constraint gap of $\{X_i\}$. Write $s_{i,\max}$ and $s^+_{i,\min}$ for the largest and smallest* positive *singular values of C_i, and put $S := \max_i s^2_{i,\max}$, $s := \min_i (s^+_{i,\min})^2$. Then*

$$\frac{1}{\sigma} \;\le\; \hat{\kappa}_{C^\top} \;\le\; \frac{S}{s} \cdot \frac{1}{\sigma},$$

with equality on the left whenever every C_i has orthonormal rows. Consequently

$$\frac{1}{\sigma} \;=\; \min\left\{\hat{\kappa}_{C^\top} : \; C_i \textit{ any matrix with } \{x : C_i x = c_i\} = X_i \textit{ for all } i\right\},$$

and the ratio $\sigma \hat{\kappa}_{C^\top}$ is unbounded above: for $X_1 = \{x \in \mathbb{R}^2 : x_1 = 0\}$, $X_2 = \{x : x_2 = 0\}$, the representation $C_1 = t\, e_1^\top$, $C_2 = e_2^\top$ has $\sigma = \frac{1}{2}$ for every $t \neq 0$ while $\hat{\kappa}_{C^\top} = 2\max\{t^2, t^{-2}\} \to \infty$ as $t \to 0$.

Proof Let P_i be the orthogonal projector onto $\operatorname{ran} C_i^\top = (X_i^0)^\perp$. Then $sP_i \preceq C_i^\top C_i \preceq SP_i$ and $\ker(C_i^\top C_i) = \ker P_i = X_i^0$, so $\frac{1}{n}\sum_i C_i^\top C_i$ and $\bar{R} := \frac{1}{n}\sum_i P_i$ have the same kernel $\bigcap_i X_i^0 = X^0$ and hence the same range $R := (X^0)^\perp$. For $u \in R$, $s\langle u, \bar{R}u\rangle \le \langle u, \frac{1}{n}\sum_i C_i^\top C_i u\rangle \le S\langle u, \bar{R}u\rangle$. Minimising over the unit sphere of R and using $\lambda^+_{\min}(\bar{R}) = \sigma$ (Proposition 3.4, since $\bar{R} = I - \bar{P}$ on R) gives $s\sigma \le \lambda^+_{\min}\big(\frac{1}{n}\sum_i C_i^\top C_i\big) \le S\sigma$. Dividing $\max_i \lambda_{\max}(C_i^\top C_i) = S$ by these two bounds gives the displayed chain. If every C_i has orthonormal rows then $C_i^\top C_i = P_i$, so $S = s = 1$ and both inequalities are equalities. Every collection admits such a representation (replace C_i by the matrix whose rows are an orthonormal basis of $\operatorname{ran} C_i^\top$, and c_i accordingly), which gives the minimum. The example is a direct computation: $\frac{1}{2}(C_1^\top C_1 + C_2^\top C_2) = \frac{1}{2}\operatorname{diag}(t^2, 1)$ and $\max_i \lambda_{\max}(C_i^\top C_i) = \max\{t^2, 1\}$, while $\frac{1}{2}(P_1 + P_2) = \frac{1}{2}I$. $\square$

Remark 5.8 So σ is what remains of $\hat{\kappa}_{C^\top}$ after every representation-dependent factor is removed, and the projection oracle attains the minimum automatically: in the projection-oracle model the representation-dependent row conditioning of the C_i is factored out by the oracle itself, and never appears in the complexity. In the affine case the gap between the two models is, admittedly, one local orthonormalization per agent, computed once offline; the lower bound of Yarmoshik et al. [35] formally excludes that step, so neither result implies the other, but we do not rest the contribution of this paper on the affine case alone. The point of σ is that it is defined, and finite, for *arbitrary* closed convex X_i, where no matrix, no singular value and no orthonormalization is available — see Section 3 and Open Problem 6.7. Proposition 5.7 was verified numerically on 200 random collections (Section 7).

6 The algorithm and its complexity

6.1 Local oracle

Agent i evaluates the *constrained conjugate gradient*

$$\mathcal{O}_i(z) := \operatorname*{arg\,min}_{x \in X_i} \{ f_i(x) + \langle z, x \rangle \} = \nabla (f_i + \iota_{X_i})^*(-z). \tag{6.1}$$

This is the constrained analogue of the oracle ∇f_i^* used by Scaman et al. [29]; for affine X_i and quadratic f_i it is one linear solve of size $\dim X_i^0$. It is a strong oracle, and we therefore report three complexities separately — communication rounds, calls to $\mathcal{O}_i$, and (in Remark 6.6) the cost of solving (6.1) numerically from ∇f_i and P_{X_i} — so that no computational cost is hidden. The lower bounds of the companion paper [12] are proved in an oracle class containing (6.1), so the comparison there is not in our favour by construction.

6.2 Accelerated gossip

Lemma 6.1 (Chebyshev gossip) *Let W satisfy Assumption 2.3 with $\gamma < 1$, put $c_\gamma := \frac{1+\gamma}{1-\gamma}$, and for $K \in \mathbb{N}$ define*

$$q_K(t) := \frac{T_K\left(\frac{2t-(1+\gamma)}{1-\gamma}\right)}{T_K(-c_\gamma)}, \qquad P_K(t) := 1 - q_K(t), \qquad \widetilde{\mathbf{W}} := P_K(\mathbf{W}).$$

Then $\widetilde{\mathbf{W}}$ is self-adjoint, positive semidefinite, satisfies $\ker \widetilde{\mathbf{W}} = \mathcal{C}$, its spectrum on $\mathcal{C}^\perp$ lies in $[1-\delta_K, 1+\delta_K]$ with $\delta_K := 1/T_K(c_\gamma)$, and one application costs exactly K communication rounds. For $K = \lceil \gamma^{-1/2} \rceil$ one has $\delta_K \le 1/\cosh 2 < 0.266$ and

$$\widetilde{\chi} := \frac{\lambda_{\max}(\widetilde{\mathbf{W}})}{\lambda_{\min}^+(\widetilde{\mathbf{W}})} \le \frac{1+\delta_K}{1-\delta_K} < 1.73.$$

If $\gamma = 1$ we set $K = 1$, $\widetilde{\mathbf{W}} = \mathbf{W}$, $\delta_K = 0$, $\widetilde{\chi} = 1$.

Proof $q_K(0) = 1$ gives $P_K(0) = 0$, hence $\ker \widetilde{\mathbf{W}} = \ker \mathbf{W} = \mathcal{C}$. For $t \in [\gamma, 1]$ the argument $\frac{2t-(1+\gamma)}{1-\gamma}$ lies in $[-1, 1]$ where $|T_K| \le 1$, and $|T_K(-c_\gamma)| = T_K(c_\gamma)$, so $|q_K(t)| \le \delta_K$ and $P_K(t) \in [1-\delta_K, 1+\delta_K]$; since $\delta_K < 1$ this is positive, so $\widetilde{\mathbf{W}} \succeq 0$. As P_K has degree K with $P_K(0) = 0$, the three-term Chebyshev recursion evaluates $P_K(\mathbf{W})x$ with K multiplications by $\mathbf{W}$. Finally $\operatorname{arccosh}(c_\gamma) = \log \frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}} \ge 2\sqrt{\gamma}$, so $T_K(c_\gamma) = \cosh(K \operatorname{arccosh} c_\gamma) \ge \cosh 2$ when $K \ge \gamma^{-1/2}$. $\square$

The construction is Chebyshev acceleration of a stationary iteration [2], introduced into decentralized optimization by Scaman et al. [29, §4.2]; we restate it in the normalization used below. Note that $\widetilde{\mathbf{W}}$ satisfies Assumption 2.3 after rescaling but *not* Assumption 2.4: it is a degree-K polynomial in $\mathbf{W}$ and

Algorithm 1 CANDA

Require: gossip matrix W with gap γ; constants μ, L; an estimate $\hat{\sigma} \in (0, \sigma]$

$K \leftarrow \lceil \gamma^{-1/2} \rceil$; $\delta \leftarrow 1/T_K(\frac{1+\gamma}{1-\gamma})$ ▷ $K \leftarrow 1,\ \delta \leftarrow 0$ if $\gamma = 1$

$\Lambda \leftarrow (1+\delta)/\mu$; $\nu \leftarrow (1-\delta)\hat{\sigma}/L$; $q \leftarrow \sqrt{\nu/\Lambda}$; $\beta \leftarrow \frac{1-q}{1+q}$

$z^0 \leftarrow 0$, $w^0 \leftarrow 0$

for $k = 0, 1, 2, \ldots$ **do**

  $s^k \leftarrow \widetilde{\mathbf{W}} w^k$ ▷ K communication rounds (Chebyshev recursion)

  $x_i^k \leftarrow \mathcal{O}_i(s_i^k)$ for each i ▷ one local oracle call

  $z^{k+1} \leftarrow w^k + \Lambda^{-1} x^k$

  $w^{k+1} \leftarrow z^{k+1} + \beta\,(z^{k+1} - z^k)$

end for

Ensure: agent i returns x_i^k

so has the sparsity of the K-th graph power. Only Assumption 2.3 is used in Section 5; Assumption 2.4 is used only to price one application of $\mathbf{W}$ at one round.

6.3 CANDA

Applying Nesterov's accelerated gradient method to Ψ built from $\widetilde{\mathbf{W}}$, and substituting $y = \widetilde{\mathbf{A}} z$ with $\widetilde{\mathbf{A}} = \widetilde{\mathbf{W}}^{1/2}$ so that only $\widetilde{\mathbf{W}}$ is ever applied, gives Algorithm 1 (CANDA: *Constraint-Aware Networked Dual Acceleration*). Every operation is local except the single call to $\widetilde{\mathbf{W}}$.

Theorem 6.2 (Complexity of CANDA) *Let Assumptions 2.1, 2.2, 2.3, 2.4 and 5.1 hold, $\mathcal{X}_0 \not\subseteq \mathcal{C}$, and let $\hat{\sigma} \in (0, \sigma]$. Then $x_i^k \in X_i$ for all i, k, and*

$$\frac{1}{n}\sum_{i=1}^{n} \|x_i^k - x^\star\|^2 \;\le\; C\Big(1 - \sqrt{\tfrac{\hat{\sigma}}{1.73\,\kappa}}\Big)^k$$

where, with $\Lambda = (1+\delta)/\mu$ and $\nu = (1-\delta)\hat{\sigma}/L$ the parameters of Algorithm 1,

$$C \;=\; \frac{c_0}{\mu^2}\Big(\frac{\Lambda}{\nu} + 1\Big)\|y^0 - y^\star\|^2, \qquad \frac{\Lambda}{\nu} \le \frac{1.73\,\kappa}{\hat{\sigma}}, \tag{6.2}$$

and $c_0 \le (1+2\beta)^2 \le 9$ an absolute constant. Consequently, to reach $\frac{1}{n}\sum_i \|x_i^k - x^\star\|^2 \le \varepsilon^2$ the algorithm uses

$$O\Big(\sqrt{\tfrac{\kappa}{\hat{\sigma}}}\,\log\tfrac{\sqrt{C}}{\varepsilon}\Big) \text{ calls to } \mathcal{O}_i \text{ per agent}, \qquad O\Big(\sqrt{\tfrac{\kappa\chi}{\hat{\sigma}}}\,\log\tfrac{\sqrt{C}}{\varepsilon}\Big) \text{ communication rounds.}$$

Remark 6.3 (On the constant, and the shorthand $\log\frac{1}{\varepsilon}$) By (6.2), C is not purely an initialization constant: it carries the factor $\Lambda/\nu = O(\kappa/\hat{\sigma})$ that every estimate-sequence bound carries, so $\log\sqrt{C} = \log\frac{1}{\varepsilon}$-independent but grows like $\frac{1}{2}\log\frac{\kappa}{\hat{\sigma}}$. The oracle-call count is therefore $O\big(\sqrt{\kappa/\hat{\sigma}}\,\log\frac{\kappa}{\hat{\sigma}\varepsilon}\big)$, i.e. $O(\sqrt{\kappa/\hat{\sigma}}\log\frac{1}{\varepsilon})$ plus an additive $O(\sqrt{\kappa/\hat{\sigma}}\log\frac{\kappa}{\hat{\sigma}})$. We write $O(\sqrt{\kappa/\hat{\sigma}}\log\frac{1}{\varepsilon})$ elsewhere in this paper, suppressing that additive term; this is the same convention under which the optimal unconstrained rate is quoted as $O(\sqrt{\kappa\chi}\log\frac{1}{\varepsilon})$ [29, Thm. 1], and the term is dominated as soon as $\varepsilon \le \hat{\sigma}/\kappa$.

Proof Feasibility $x_i^k \in X_i$ is immediate from (6.1). Replace $\mathbf{W}$ by $\widetilde{\mathbf{W}}$ throughout Section 5. By Lemma 6.1, $\widetilde{\mathbf{W}}/\lambda_{\max}(\widetilde{\mathbf{W}})$ satisfies Assumption 2.3 with gossip condition number $\widetilde{\chi} < 1.73$, and $\ker \widetilde{\mathbf{W}} = \mathcal{C}$ so the constraint encoded is unchanged. Applying Theorem 5.3 with $\lambda_{\max}(\widetilde{\mathbf{W}}) \le 1+\delta$ and $\lambda^+_{\min}(\widetilde{\mathbf{W}}) \ge 1-\delta$ gives that Ψ is $\frac{1+\delta}{\mu}$-smooth and $\frac{(1-\delta)\sigma}{L}$-strongly convex on $\mathcal{Y}$; the values Λ and ν of Algorithm 1 are exactly these two constants with σ replaced by the underestimate $\hat{\sigma}$, which only decreases the assumed strong convexity, so $\kappa_{\text{dual}} \le \Lambda/\nu \le 1.73\,\kappa/\hat{\sigma}$.

By Theorem 5.3(ii) the gradient of Ψ lies in $\mathcal{Y}$, so a trajectory started at $y^0 = 0$ remains in $\mathcal{Y}$; there Nesterov's constant-step scheme for smooth strongly convex functions [23, Thm. 2.2.3] gives $\Psi(y^k) - \Psi^\star \le (1 - \sqrt{\nu/\Lambda})^k\big(\Psi(y^0) - \Psi^\star + \frac{\nu}{2}\|y^0 - y^\star\|^2\big)$ and hence $\|y^k - y^\star\|^2 \le \frac{2}{\nu}(\Psi(y^k) - \Psi^\star)$ decays at the same rate, where $y^\star \in \mathcal{Y}$ may be chosen because Ψ is constant along $\ker B$. Since $\nabla\Psi(y^\star) = 0$ forces $x(y^\star) \in \mathcal{X} \cap \ker \mathbf{A} = \mathcal{X} \cap \mathcal{C}$ and $\langle \mathbf{A}y^\star, x\rangle = 0$ there, $x(y^\star)$ minimizes F over $\mathcal{X} \cap \mathcal{C}$, i.e. $x(y^\star) = \mathrm{J}(x^\star)$. Two book-keeping points complete the estimate. First, Algorithm 1 evaluates the oracle at the *extrapolated* point v^k, not at y^k; since $v^k = y^k + \beta(y^k - y^{k-1})$ with $\beta \in [0, 1)$, we have $\|v^k - y^\star\| \le (1 + 2\beta)\max\{\|y^k - y^\star\|, \|y^{k-1} - y^\star\|\}$, which costs the absolute factor $c_0 = (1 + 2\beta)^2 \le 9$ in (6.2) and nothing in the rate. Second, after replacing $\mathbf{W}$ by $\widetilde{\mathbf{W}}$ the last claim of Theorem 5.3 reads $\|x(y) - x(y')\| \le \frac{\|B\|}{\mu}\|y - y'\|$ with $\|B\| \le \lambda_{\max}(\widetilde{\mathbf{W}})^{1/2} \le \sqrt{1+\delta} \le 1.09$, which we absorb into c_0. Combining with $\|y^k - y^\star\|^2 \le \frac{2}{\nu}(\Psi(y^k) - \Psi^\star)$ and $\Psi(y^0) - \Psi^\star \le \frac{\Lambda}{2}\|y^0 - y^\star\|^2$ gives (6.2) and the displayed bound in the averaged norm.

Finally the substitution $y = \widetilde{\mathbf{A}}z$, $v = \widetilde{\mathbf{A}}w$ turns $y^{k+1} = v^k - \Lambda^{-1}\nabla\Psi(v^k)$ with $\nabla\Psi(v) = -\widetilde{\mathbf{A}}x(v)$ and $x(\widetilde{\mathbf{A}}w) = \arg\min_{x\in\mathcal{X}}\{F(x) + \langle \widetilde{\mathbf{W}}w, x\rangle\}$ into $z^{k+1} = w^k + \Lambda^{-1}x^k$, the extrapolation transferring verbatim; and F, $\mathcal{X}$ are separable, so the argmin decomposes into the n local problems (6.1). Each iteration costs one oracle call and $K \le \gamma^{-1/2} + 1$ rounds. $\square$

Remark 6.4 (Per-agent constants) Because $\|\cdot\|$ is the averaged norm, the per-agent statement carries a factor n: $\|x_i^k - x^\star\|^2 \le nC(1 - \sqrt{\hat{\sigma}/(1.73\kappa)})^k$. The iterates satisfy $x_i^k \in X_i$ but are neither consensual nor in X; combining Definition 3.1 with $\operatorname{dist}(x_i^k, X_i) = 0$ gives $\operatorname{dist}^2(\bar{x}^k, X) \le \sigma^{-1}\frac{1}{n}\sum_i\|x_i^k - \bar{x}^k\|^2$, so the feasibility residual of the average decays at the same linear rate.

Remark 6.5 (Consistency checks) With $\sigma = 1$ (all $X_i = \mathbb{R}^d$) Theorem 6.2 gives $O(\sqrt{\kappa}\log\frac{1}{\varepsilon})$ oracle calls and $O(\sqrt{\kappa\chi}\log\frac{1}{\varepsilon})$ communications, optimal by Scaman et al. [29]. With $\kappa = \chi = 1$ it gives $O(\sigma^{-1/2}\log\frac{1}{\varepsilon})$, which by Proposition 3.9(b) is $O(\sin^{-1}\theta_F \log\frac{1}{\varepsilon})$ for two sets — the rate of optimally tuned generalized alternating projections [9], and order-optimal by the two-set corollary of [12].

Remark 6.6 (Primal oracle) If agent i can only evaluate ∇f_i and P_{X_i}, the subproblem (6.1) can be solved by accelerated projected gradient in $O(\sqrt{\kappa}\log\frac{1}{\eta})$

steps. Turning this into a complete complexity statement requires an inexactness model for accelerated methods — the (δ, L)-oracle of Devolder et al. [8] shows that error does not in general vanish under acceleration, so a summability schedule $\eta_k \downarrow 0$ in the sense of Schmidt et al. [30] is needed. We do not carry this out and do not claim the resulting bound; establishing whether $O(\sqrt{\kappa/\sigma})$ projections suffice is the projection-oracle analogue of the gradient/communication separation studied for affine coupling by Yarmoshik et al. [36].

6.4 Obtaining σ

Algorithm 1 needs $\hat{\sigma}$, and σ encodes the joint geometry of all the private sets — exactly the global object the decentralized setting forbids any agent from forming. Three points make this workable.

First, the guarantee is monotone: Theorem 6.2 holds for *every* $\hat{\sigma} \in (0, \sigma]$, at the price of a factor $\sqrt{\sigma/\hat{\sigma}}$ in the iteration count. Underestimating is safe; overestimating is not covered.

Second, σ can be *approached* by a power iteration on the simultaneous projection Q of Proposition 3.7: the ratio $\|z^{k+1} - z^k\| / \|z^k - z^{k-1}\|$ converges to $\lambda_{\max}(\bar{P}|_{(X^0)^\perp}) = 1 - \sigma$. This is a *heuristic*, not a certified procedure, for three reasons, and we state them plainly. *(i)* The ratio $\operatorname{dist}(Q(z), X)/\operatorname{dist}(z, X)$ must not be used instead: $\operatorname{dist}(\cdot, X)$ involves the global feasible set, which is precisely what no agent can form. *(ii)* Applying Q even once is not a local operation: $Q(z) = \frac{1}{n}\sum_i P_{X_i}(z)$ averages over *all* agents, so on a sparse network each power iteration requires an inner average-consensus procedure, not one projection step per agent. The same is true of the ratio itself, which involves the global norm $\|z\|^2 = \frac{1}{n}\sum_i \|z_i\|^2$; a single sparse gossip round evaluates neither, and any decentralized average-consensus subroutine returns them only up to an error that decays geometrically in its own inner iterations. The procedure is therefore an offline preprocessing idea, not a subroutine of Algorithm 1. *(iii)* Even with exact norms the estimate is one-sided the wrong way: Rayleigh quotients approach $\lambda_{\max}$ from below, so 1 minus the observed ratio approaches σ from *above*, and every finite-k value is an overestimate — the direction Theorem 6.2 does not cover. A deflation factor must therefore be applied, and we have no finite-k guarantee for how large it must be. We regard this as a genuine gap rather than a detail; σ, like γ, is a global parameter of the instance, and neither is computed by the algorithm.

Third, in the polyhedral case a valid $\hat{\sigma}$ is often available in closed form: Proposition 3.9(a) gives $\sigma = \lambda_{\min}^{+}(\frac{1}{n}\sum_i a_i a_i^\top)$ for hyperplanes, so any lower bound on that smallest *positive* eigenvalue is admissible. Note that a Gershgorin estimate does not serve here: it bounds $\lambda_{\min}$, which is 0 whenever X is not a single point. In practice $\hat{\sigma}$ should come from the structure of the application, and we regard supplying it as part of the modelling rather than of the algorithm.

6.5 Beyond affine local sets

Assumption 5.1 enters only through the affine structure of $\mathcal{X}$, but it does so in more than one step of the proof of Theorem 5.3: the conjugate identity $G^*(z) = \langle z, \mathrm{J}(x^0)\rangle + \varphi^*(E^*z)$, which requires $\mathcal{X}$ to be a translate of a subspace; the reduction $\lambda^+_{\min}(R) = 1 - c_F(\mathcal{X}_0^\perp, \mathcal{C}^\perp)^2 = \sigma$, which goes through Proposition 3.4 and so presupposes that the direction spaces X_i^0 exist at all; and the strong-duality step $\operatorname{ri}\mathcal{X} = \mathcal{X}$. The companion paper [12] uses it once more, in the form that $P_{\mathcal{X}}$ is affine with linear part $P_{\mathcal{X}_0}$. Section 4 does not use it, and Section 3 uses it only where marked.

For general closed convex X_i the conjugate G^* need not be differentiable and the argument breaks down. Two routes are natural and we leave both open. For *polyhedral* X_i and *quadratic* f_i, G^* is piecewise quadratic and one expects a global linear rate through a Luo–Tseng-type error bound, with σ replaced by the gap of the affine hulls of the active faces at the solution. For general μ-strongly convex, L-smooth f_i this is false — restricting a smooth strongly convex function to a polyhedron does not make its conjugate piecewise quadratic — and the analogous argument would instead require a metric-subregularity or error-bound property of the dual gradient mapping, which we do not establish. For general convex X_i one expects a local result under partial smoothness [19]: the relevant nondegeneracy condition is on the *shifted* subproblem actually solved by CANDA, namely $-\big(\nabla f_i(x^\star) + (\mathbf{A}y^\star)_i\big) \in \operatorname{ri} N_{X_i}(x^\star)$, and identification of the active manifolds would follow from Hare and Lewis [10, Thm. 5.3]. We emphasise that identification alone is not sufficient: after identification the problem is posed on a curved manifold $\mathcal{M}_i$, and replacing $\mathcal{M}_i$ by its tangent affine space is a first-order approximation whose curvature contribution must be controlled. We do not have that argument, and state the matter as an open problem rather than a theorem.

Open Problem 6.7 Establish a global linear rate for (P) with polyhedral X_i, and a local one for general closed convex X_i, with the constant expressed through the constraint gap of the corresponding tangent structures.

7 Numerical illustration

The role of this section is to test the theory, not to argue practical superiority. All code, seeds and scripts are in the repository accompanying this paper; every number below is reproduced by a single script.

Instances. n agents, $x \in \mathbb{R}^d$, $f_i(x) = \frac{1}{2}\langle x, H_i x\rangle - \langle g_i, x\rangle$ with $\operatorname{spec}(H_i) \subseteq [\mu, L]$, $\mu = 1$, and $X_i = \{x : \langle a_i, x\rangle = b_i\}$ a private hyperplane. Normals are $a_i \propto a_0 + \varepsilon u_i$ as in Example 3.10, so $\sigma = \lambda^+_{\min}(\frac{1}{n}\sum_i a_i a_i^\top)$ is tuned by ε while each individual constraint stays perfectly conditioned; the normals lie in a fixed coordinate subspace of the shared eigenbasis of the H_i, so the

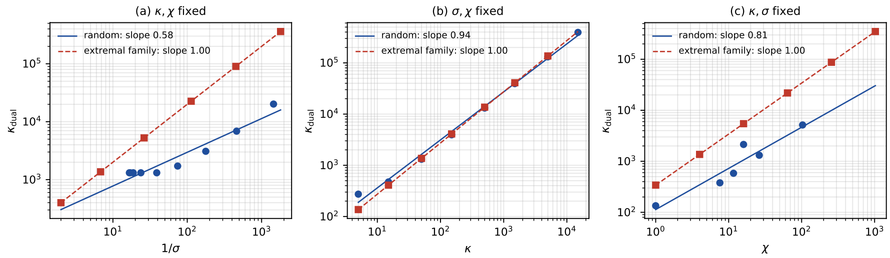

Fig. 1 Dual condition number against each parameter with the other two fixed, for random instances (circles) and for the extremal family of Theorem 5.5 (squares). The bound $\kappa_{\mathrm{dual}} \leq \kappa\chi/\sigma$ predicts slope at most 1, with equality on the extremal family.

problem restricted to X has condition number exactly κ. The gossip matrix is the combinatorial graph Laplacian normalized by its largest eigenvalue. Unless stated otherwise $n = 16$, $d = 24$, $\dim X = 18$, $\kappa = 50$, ring graph ($\chi = 26.3$), seed 4.

Proposition 3.4. Computing σ from Definition 3.1 via $\lambda^+_{\min}(\frac{1}{n}\sum a_i a_i^\top)$ and independently as $1 - c_F(\mathcal{X}_0, \mathcal{C})^2$ by a principal-angle computation in $\mathbb{R}^{nd}$ agrees to machine precision (4.2×10^{-17} absolute on the reported instance). The fourth identity of Proposition 3.4, $\sigma = \frac{n-1}{n}(1 - c_{\mathrm{RZ}})$, was checked on 40 random collections of subspaces with $n \in \{2, \ldots, 5\}$ in $\mathbb{R}^7$; the largest absolute discrepancy was 1.1×10^{-15}. This validates the implementation, not the propositions.

Proposition 5.7. On 200 random affine collections ($n \in \{2, \ldots, 5\}$, $d = 8$, 1–3 rows per agent, half of them row-orthonormalized and a third randomly rescaled per agent) the chain $1/\sigma \leq \hat{\kappa}_{C^\top} \leq (S/s)/\sigma$ held in every instance. On 200 row-orthonormal representations the maximum relative deviation of $\hat{\kappa}_{C^\top}$ from $1/\sigma$ was 6.8×10^{-11}. On the two-agent example of Proposition 5.7, σ stays at $\frac{1}{2}$ while $\hat{\kappa}_{C^\top}$ runs from 2 to 2×10^6 as t falls from 1 to 10^{-3}.

Theorems 5.3 and 5.5. On 27 random instances spanning three graph families, $n \in \{6, 10, 14\}$ and $\kappa \in \{5, 50, 500\}$, the bound $\kappa_{\mathrm{dual}} \leq \kappa\chi/\sigma$ held in every case, with ratio in $[0.124, 0.291]$ (and 0.018 on the larger instance of the next paragraph); the bound is therefore conservative on random data. On the extremal family of Theorem 5.5, 81 instances spanning $n \in \{4, 6, 10\}$, $\kappa \in \{1, 7, 60\}$, $\chi \in \{1, 4, 25\}$ and three angles reproduce $\kappa_{\mathrm{dual}} = \kappa\chi/\sigma$ to within 1.4×10^{-12} relative. Figure 1 shows the two behaviours side by side: the fitted log–log slopes of κ_{dual} are $0.58, 0.94, 0.81$ in $\sigma^{-1}, \kappa, \chi$ on random instances — all strictly below the theoretical exponent 1, as the inequality permits — and $1.00, 1.00, 1.00$ on the extremal family, where Theorem 5.5 predicts equality.

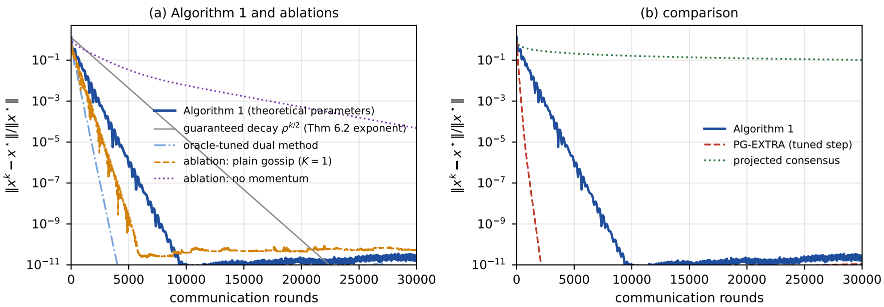

Fig. 2 (a) Algorithm 1 with theoretical parameters, its guaranteed rate, an oracle-tuned reference, and two ablations. (b) Comparison with PG-EXTRA and projected consensus.

Theorem 6.2. Figure 2(a) runs Algorithm 1 *exactly as printed*, with Λ and m computed from $(\mu, L, \gamma, \hat{\sigma} = \sigma)$ — global parameters of the instance, but no information whatever about the spectrum of the dual Hessian — against the rate it is guaranteed. On the reported instance ($\kappa = 50$, $\chi = 26.3$, $\sigma = 1.35 \times 10^{-2}$, $K = 6$, $\delta_K = 0.185$) it reaches relative error 10^{-8} in 6702 communication rounds. The reference curve in Figure 2(a) plots the *exponent* guaranteed by Theorem 6.2, $\rho^{k/2}$ with $\rho = 1 - \sqrt{\nu/\Lambda}$ — the square root because the theorem bounds the squared averaged error — anchored at the observed initial error rather than at $\sqrt{C}$; so it is optimistic in the constant and honest in the rate. Even so it reaches 10^{-8} only at 16 365 rounds, against the measured 6702. For reference we also plot an oracle-tuned variant, which uses the exact extreme eigenvalues of the dual Hessian (a centralized computation, not an implementable algorithm) and needs 2892 rounds; the gap is the price of the conservative constants in Theorem 5.3: the operator Algorithm 1 actually runs on is the Chebyshev-preconditioned dual, whose exact condition number is $\kappa_{\text{dual}}(\widetilde{\mathbf{W}}) = 6.96 \times 10^2$, against the bound $1.73\kappa/\sigma = 6.4 \times 10^3$ — an overestimate by a factor 9.2. (For the raw gossip operator the corresponding figures are $\kappa_{\text{dual}}(\mathbf{W}) = 1.7 \times 10^3$ against $\kappa\chi/\sigma = 9.8 \times 10^4$.) Across five seeds Algorithm 1 needs 5622–6702 rounds.

Two ablations. Removing the momentum ($\beta = 0$) prevents convergence to 10^{-8} within a budget of 3×10^4 rounds, confirming that acceleration is doing the work. Removing the Chebyshev step ($K = 1$) *improves* communication, to 4078 rounds. This is not a contradiction: as Figure 3 shows, Chebyshev gossip does not reduce communication — Theorem 6.2 gives both variants the same $O(\sqrt{\kappa\chi/\sigma})$ round bound and the polynomial pays a constant on top — but it removes the network entirely from the *computational* cost. The measured log–log slope of local oracle calls against χ is 0.02 with Chebyshev gossip and 0.48 without. Theorem 6.2 predicts *upper* bounds with slopes 0 and $\frac{1}{2}$ respectively (for $K = 1$ the same proof gives $\nu = \hat{\sigma}\gamma/L$), so the measurements are consistent with it; we do not have a matching lower bound for either variant, so this is corroboration rather than confirmation.

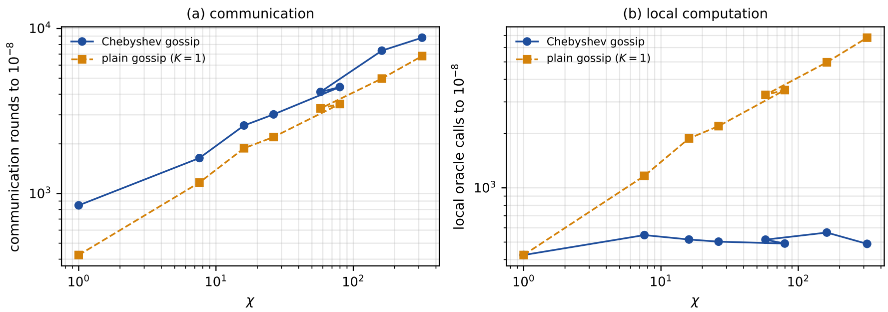

Fig. 3 Chebyshev gossip does not reduce communication rounds (a) — plain gossip is a constant factor better across the whole range tested — but it makes the number of local oracle calls independent of the network (b): the fitted log–log slopes in χ of the oracle-call counts are 0.02 with Chebyshev gossip and 0.48 without, consistent with the upper bounds of Theorem 6.2, whose exponents are 0 and $\frac{1}{2}$.

Sensitivity to $\hat{\sigma}$. Theorem 6.2 allows any $\hat{\sigma} \leq \sigma$ at a cost $\sqrt{\sigma/\hat{\sigma}}$. Measured: $\hat{\sigma} = \sigma/2, \sigma/5, \sigma/20$ give 8946, 12852, 18756 rounds against 6702 at $\hat{\sigma} = \sigma$, i.e. factors $1.33, 1.92, 2.80$ where the theory permits $1.41, 2.24, 4.47$. Mild overestimation was benign on these instances ($\hat{\sigma} = 2\sigma$ and 5σ converged, in 5058 and 3300 rounds) but is not covered by the theorem.

Comparison. Figure 2(b) compares with PG-EXTRA [32] using $r_i = \iota_{X_i}$ and a step size selected by offline grid search, and with projected consensus [22]. Projected consensus is visibly sublinear and does not reach 10^{-8} within the budget. Tuned PG-EXTRA reaches it in 1390 rounds, faster than Algorithm 1, and we report this plainly. The companion paper [12] explains why: PG-EXTRA is itself a second-order recursion driven by the same operator, and empirically attains the $\sqrt{\kappa\chi/\sigma}$ scaling, so the difference is a constant and not an exponent. Two further remarks. PG-EXTRA carries no linear-rate guarantee for heterogeneous r_i — by Alghunaim et al. [1, Thm. 3], *without an additional regularity assumption* no method in that oracle class admits a worst-case global linear guarantee — so its behaviour here is a property of this instance family, whereas Theorem 6.2 certifies Algorithm 1 on every affine instance with $\sigma > 0$. And PG-EXTRA's step size was chosen with knowledge of the solution, while Algorithm 1 used only $(\mu, L, \gamma, \hat{\sigma})$. We regard closing the constant-factor gap — most plausibly by an adaptive restart that estimates κ_{dual} online rather than bounding it — as the main practical question left open.

8 Concluding remarks

Decentralized optimization over an intersection of private convex sets is governed by two spectral gaps of the same kind. One, γ, belongs to the communication graph and measures how quickly local averaging enforces agreement. The other, σ, belongs to the constraint geometry and measures how quickly local projections enforce feasibility; for affine sets it is the spectral gap of the simultaneous projection operator, a reading we take from Reich and Zalas [27]. Once σ is in place the qualitative obstruction to linear convergence disappears: optimality certificates separate across agents (Theorem 4.1), the solution map is stable (Theorem 4.2), the dual conditioning is at most $\kappa\chi/\sigma$ and no better (Theorems 5.3 and 5.5), and an accelerated dual method attains $O(\sqrt{\kappa\chi/\sigma}\log\frac{1}{\varepsilon})$ communication rounds and $O(\sqrt{\kappa/\sigma}\log\frac{1}{\varepsilon})$ local oracle calls (Theorem 6.2). The exponent $\frac{1}{2}$ in σ cannot be improved, and the two-set corollary settles the order-optimality of generalized alternating projections; both are proved in the companion paper [12].

That σ belongs to the problem and not to our method is, we think, the point worth keeping, and it is the subject of the companion paper [12]: the operator behind Theorem 5.3 is the same one behind PG-EXTRA, a primal method designed without any reference to constraint regularity, and no step size makes PG-EXTRA faster than $\sqrt{(\kappa-1)\chi/\sigma}$ on the extremal family. What an analysis buys is therefore not speed but certainty: Algorithm 1 attains $O(\sqrt{\kappa\chi/\sigma}\log\frac{1}{\varepsilon})$ from μ, L, γ and any underestimate of σ, on every affine instance with $\sigma > 0$.

Three questions remain, stated above as Open Problem 6.7, in Remark 6.6, and in [12]: a global rate beyond affine local sets, the multiplicative lower bound in the full oracle class, and a complexity statement for the weaker $(\nabla f_i, P_{X_i})$ oracle. It would also be natural to ask how σ behaves under time-varying or stochastic constraints and on directed graphs, where the analogue of Chebyshev acceleration is known to fail [14].

Declarations

Funding. Not applicable.

Competing interests. The author declares no competing interests.

Data and code availability. All code, random seeds and scripts reproducing every figure and every numerical claim in this paper are archived on Zenodo at `10.5281/zenodo.21961115` [13], together with a `README` mapping each script to the statement it reproduces. No external data are used: every instance is generated from a seeded pseudorandom stream by the included code.

Author contributions. Sole author: conception, analysis, software and writing.

Use of AI tools. The author used a large language model (Claude, Anthropic) as an assistant throughout the preparation of this manuscript, including drafting and revising text, developing and checking proofs, writing the numerical verification code, and searching the literature. All mathematical statements, proofs and numerical claims were reviewed, tested and verified by the author, who takes full responsibility for the content of this article.

References

1. Alghunaim, S.A., Ryu, E.K., Yuan, K., Sayed, A.H.: Decentralized proximal gradient algorithms with linear convergence rates. IEEE Transactions on Automatic Control **66**(6), 2787–2794 (2021). DOI 10.1109/TAC.2020.3009363
2. Arioli, M., Scott, J.: Chebyshev acceleration of iterative refinement. Numerical Algorithms **66**(3), 591–608 (2014)
3. Bauschke, H.H., Borwein, J.M.: On projection algorithms for solving convex feasibility problems. SIAM Review **38**(3), 367–426 (1996). DOI 10.1137/S0036144593251710
4. Bauschke, H.H., Borwein, J.M., Li, W.: 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). DOI 10.1007/s101070050083
5. Dao, M.N., Dizon, N.D., Hogan, J.A., Tam, M.K.: Constraint reduction reformulations for projection algorithms with applications to wavelet construction. Journal of Optimization Theory and Applications **190**(1), 201–233 (2021). DOI 10.1007/s10957-021-01878-z
6. Deutsch, F.: The angle between subspaces of a Hilbert space. In: S.P. Singh (ed.) Approximation Theory, Wavelets and Applications, *NATO ASI Series C: Mathematical and Physical Sciences*, vol. 454, pp. 107–130. Kluwer (1995). DOI 10.1007/978-94-015-8577-4_7
7. Deutsch, F.: Best Approximation in Inner Product Spaces. CMS Books in Mathematics. Springer, New York (2001)
8. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. Mathematical Programming **146**(1–2), 37–75 (2014). DOI 10.1007/s10107-013-0677-5
9. Fält, M., Giselsson, P.: Optimal convergence rates for generalized alternating projections. In: 56th IEEE Conference on Decision and Control (CDC), pp. 2268–2274 (2017). DOI 10.1109/CDC.2017.8263980. ArXiv:1703.10547
10. Hare, W.L., Lewis, A.S.: Identifying active constraints via partial smoothness and prox-regularity. Journal of Convex Analysis **11**(2), 251–266 (2004)
11. Hoffman, A.J.: On approximate solutions of systems of linear inequalities. Journal of Research of the National Bureau of Standards **49**(4), 263–265 (1952)
12. Jha, S.: Constraint-Gap Lower Bounds for Decentralized Optimization with Private Constraints. Companion manuscript (2026). Companion to the present paper
13. Jha, S.: Reproduction Code for "Two Spectral Gaps" and "Constraint-Gap Lower Bounds": Decentralized Optimization over Intersections of Local Convex Sets. Zenodo (2026). DOI 10.5281/zenodo.21961115. URL https://doi.org/10.5281/zenodo.21961115

14. Kovalev, D., Gasanov, E., Gasnikov, A., Richtárik, P.: Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. In: Advances in Neural Information Processing Systems 34 (NeurIPS) (2021)
15. Kruger, A.Y., Luke, D.R., Thao, N.H.: Set regularities and feasibility problems. Mathematical Programming **168**(1–2), 279–311 (2018). DOI 10.1007/s10107-016-1039-x
16. Kruger, A.Y., Thao, N.H.: Quantitative characterizations of regularity properties of collections of sets. Journal of Optimization Theory and Applications **164**(1), 41–67 (2015). DOI 10.1007/s10957-014-0556-0
17. Lee, S., Nedić, A.: Distributed random projection algorithm for convex optimization. IEEE Journal of Selected Topics in Signal Processing **7**(2), 221–229 (2013). DOI 10.1109/JSTSP.2013.2247023
18. Lee, S., Nedić, A.: Asynchronous gossip-based random projection algorithms over networks. IEEE Transactions on Automatic Control **61**(4), 953–968 (2016). DOI 10.1109/TAC.2015.2460051
19. Lewis, A.S.: Active sets, nonsmoothness, and sensitivity. SIAM Journal on Optimization **13**(3), 702–725 (2002). DOI 10.1137/S1052623401387623
20. Mai, V.S., Abed, E.H.: Distributed optimization over directed graphs with row stochasticity and constraint regularity. Automatica **102**, 94–104 (2019). DOI 10.1016/j.automatica.2018.07.020
21. Necoara, I., Richtárik, P., Pătraşcu, A.: Randomized projection methods for convex feasibility: Conditioning and convergence rates. SIAM Journal on Optimization **29**(4), 2814–2852 (2019). DOI 10.1137/18M1167061
22. Nedić, A., Ozdaglar, A., Parrilo, P.A.: Constrained consensus and optimization in multi-agent networks. IEEE Transactions on Automatic Control **55**(4), 922–938 (2010). DOI 10.1109/TAC.2010.2041686
23. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Boston (2004). Theorem numbering as in this edition; the 2nd edition appeared as *Lectures on Convex Optimization*, Springer, 2018
24. Pang, C.H.J.: Linear convergence of distributed Dykstra's algorithm for sets under an intersection property (2018)
25. Pierra, G.: Decomposition through formalization in a product space. Mathematical Programming **28**(1), 96–115 (1984)
26. Ram, S.S., Nedić, A., Veeravalli, V.V.: Distributed stochastic subgradient projection algorithms for convex optimization. Journal of Optimization Theory and Applications **147**(3), 516–545 (2010). DOI 10.1007/s10957-010-9737-7
27. Reich, S., Zalas, R.: The optimal error bound for the method of simultaneous projections. Journal of Approximation Theory **223**, 96–107 (2017). DOI 10.1016/j.jat.2017.08.005
28. Reich, S., Zalas, R.: Error bounds for the method of simultaneous projections with infinitely many subspaces. Journal of Approximation Theory **272**, 105,648 (2021). DOI 10.1016/j.jat.2021.105648

29. Scaman, K., Bach, F., Bubeck, S., Lee, Y.T., Massoulié, L.: Optimal algorithms for smooth and strongly convex distributed optimization in networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), *PMLR*, vol. 70, pp. 3027–3036 (2017)
30. Schmidt, M., Le Roux, N., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Advances in Neural Information Processing Systems 24 (NeurIPS) (2011)
31. Shahriari-Mehr, F., Panahi, A.: Double averaging and gradient projection: Convergence guarantees for decentralized constrained optimization (2022)
32. Shi, W., Ling, Q., Wu, G., Yin, W.: A proximal gradient algorithm for decentralized composite optimization. IEEE Transactions on Signal Processing **63**(22), 6013–6023 (2015). DOI 10.1109/TSP.2015.2461520
33. Wu, X., Lu, J.: Fenchel dual gradient methods for distributed convex optimization over time-varying networks. IEEE Transactions on Automatic Control **64**(11), 4629–4636 (2019). DOI 10.1109/TAC.2019.2901829
34. Xu, Q., Aybat, N.S., Gürbüzbalaban, M.: An accelerated primal-dual algorithm with backtracking for decentralized constrained optimization (2025)
35. Yarmoshik, D., Nguyen, N.T., Rogozin, A., Gasnikov, A.: Complexity of decentralized optimization with mixed affine constraints. In: International Conference on Machine Learning (ICML), PMLR (2026). ArXiv:2602.04479
36. Yarmoshik, D., Rogozin, A., Kiselev, A., Dorin, N., Gasnikov, A., Kovalev, D.: Decentralized optimization with coupled constraints. In: International Conference on Learning Representations (ICLR) (2025). ArXiv:2407.02020
37. Yazdandoost Hamedani, E., Aybat, N.S.: A decentralized primal-dual method for constrained minimization of a strongly convex function. IEEE Transactions on Automatic Control **67**(11), 5682–5697 (2022). DOI 10.1109/TAC.2021.3130082