Computing Lewis weights to high precision using local relative smoothness

Sander Gribling S.J.GRIBLING@TILBURGUNIVERSITY.EDU
Tilburg University
Aaron Sidford SIDFORD@STANFORD.EDU
Stanford University
Chenyi Zhang CHENYIZ@STANFORD.EDU
Stanford University

Editors: Steve Hanneke and Tor Lattimore

Abstract

We provide algorithms that compute ε-estimates of the ℓ_p-Lewis weights of a matrix $A \in \mathbb{R}^{m\times n}$ for $p \geq 4$ using $O(p^2 \log(m/\varepsilon))$ rounds of leverage score computation, where ℓ_p-Lewis weights and leverage scores are both standard measures of row importance. This improves upon the state-of-the-art round complexity of $O(p^3 \log(m/\varepsilon))$ due to Fazel, Lee, Padmanabha, and Sidford (2022). We obtain our results by carefully applying a local variant of relatively smooth gradient descent to primal and dual forms of the ℓ_p-Lewis weight optimization problem and providing tools to convert between different notions of approximate ℓ_p-Lewis weights.

Keywords: ℓ_p-Lewis weights, leverage scores, relative smoothness, optimal design

1. Introduction

The *ℓ_p-Lewis weights of a matrix $A \in \mathbb{R}^{m\times n}$* , denoted $\sigma_p(A) \in \mathbb{R}^m_{\geq 0}$, are a fundamental measure of the importance of the rows of A (Lewis, 1978; Bourgain et al., 1989; Cohen and Peng, 2015). They arise in sampling schemes for sparsifying a matrix with respect to the ℓ_p-norm (Cohen and Peng, 2015; Lee, 2016; Jambulapati et al., 2022), self-concordant barriers for linear programming (Lee and Sidford, 2019; van den Brand et al., 2021), optimal design problems in statistics, and geometric problems (Todd, 2016; Cohen and Peng, 2015; Fazel et al., 2022).

ℓ_p-Lewis weights can be viewed as an ℓ_p-generalization of the *leverage scores of* A, denoted $\sigma(A)$, a natural measure of the importance of a row in ℓ_2. Leverage scores have many applications in statistics (Rudelson and Vershynin, 2007), randomized linear algebra (Drineas et al., 2012), and graph algorithms (Spielman and Srivastava, 2008). In the case that A has full (row-)rank (which we assume for simplicity), $\sigma(A)$ is defined as

$$\sigma(A)_i := a_i^\top \left(A^\top A\right)^{-1} a_i \text{ and } a_i^\top \text{ is the } i\text{'th row of } A \text{ for all } i \in [m]\,.$$

The ℓ_p-Lewis weights of A are the leverage scores of A when $p = 2$. Otherwise, they are defined implicitly as the leverage scores after appropriate re-weighting the rows by these same leverage scores. Below we define them formally for what we call *non-degenerate* matrices.[1]

. This work subsumes the note "On computing approximate Lewis weights" by Apers et al. (2024).

1. Restricting to non-degenerate matrices is not a strong assumption: zero rows have both leverage scores and Lewis weights equal to zero and a rank-deficiency can either be removed by carefully restricting to a smaller subspace or by replacing matrix inverses with pseudoinverses.

Definition 1 (ℓ_p-Lewis weights) *For $p \in (0,\infty)$ the ℓ_p-Lewis weights of non-degenerate, i.e., full-rank with no zero rows, $A \in \mathbb{R}^{m\times n}$, denoted $\sigma_p(A) \in \mathbb{R}^m_{>0}$, is the unique (Lewis, 1978; Wojtaszczyk, 1991; Cohen and Peng, 2015) positive $w \in \mathbb{R}^m_{>0}$ where, for $W = \mathrm{Diag}(w)$,*

$$w = \sigma(W^{\frac{1}{2}-\frac{1}{p}}A). \tag{1}$$

To further motivate ℓ_p-Lewis weights, we consider two fundamental problems from statistics and geometry associated with a set of vectors $\{a_i\}_{i\in[m]} \subseteq \mathbb{R}^n$, see e.g., Khachiyan (1996); Todd (2016).[2] First, the D-optimal design problem in statistics associates the $\{a_i\}_{i\in[m]}$ to experiments and asks to assign probabilities $\lambda \in \mathbb{R}^m_{\geq 0}$ to them in order to minimize the determinant of the error covariance matrix $(\sum_{i\in[m]} \lambda_i a_i a_i^\top)^{-1}$, i.e.,

$$\min_{\lambda\in\mathbb{R}^m_{\geq 0}} -\mathrm{logdet}\Big(\sum_{i\in[m]} \lambda_i a_i a_i^\top\Big) \quad \text{s.t.} \quad \mathbf{1}^\top\lambda = 1. \tag{2}$$

This corresponds to designing the experiment to minimize the volume of the resulting confidence ellipsoid (for any fixed confidence level). The (Lagrange) dual problem,

$$\min_{M\in\mathcal{S}^n_{\succ 0}} -\mathrm{logdet}(M) \quad \text{s.t.} \quad a_i^\top M a_i \leq 1 \ \ \forall i \in [m]. \tag{3}$$

is a fundamental geometric problem that computes the minimum volume ellipsoid $\mathcal{E} := \{x \in \mathbb{R}^n : x^\top M x \leq 1\}$ that contains the vectors $\{a_i\}_{i\in[m]}$, i.e., the John ellipsoid of the set $\{a_i\}_{i\in[m]}$ (Fritz, 1948). Cohen and Peng (2015) showed that an ℓ_p-variant of (3) is connected to Lewis weights: based on Wojtaszczyk (1991), they introduced the convex program

$$\min_{M\succeq 0} \det(M)^{-1} \quad \text{such that} \quad \sum_{i\in[m]} (a_i^\top M a_i)^{p/2} \leq 1\,. \tag{4}$$

The optimal solution M_* of (4) satisfies $M_*^{-1} = \sum_{i\in[m]} (a_i^\top M_* a_i)^{p/2-1} a_i a_i^\top$ and therefore encodes the Lewis weights of A as $\sigma_p(A)_i = (a_i^\top M_* a_i)^{p/2}$.

Given their applications and connections, computing ℓ_p-Lewis weights is a prominent structured optimization problem. Additionally, given the well-studied nature of leverage scores and their simple expression linear-algebraically, previous work studied how much more algorithmically challenging it is to compute Lewis weights (Cohen and Peng, 2015; Lee, 2016; Lee and Sidford, 2019; Fazel et al., 2022). Similarly, in this paper we study the following central question:

How many leverage score computations, i.e., computing $\sigma(DA)$ for diagonal D, suffice to estimate $\sigma_p(A)$?

We seek new algorithmic and analytic tools for answering this question.

1.1. Prior work

For $p \in (0,4)$, it is straightforward to show the map $m(w) := \sigma(\mathrm{Diag}(w)^{\frac{1}{2}-\frac{1}{p}}A)$ is multiplicatively contractive for $w \in \mathbb{R}^m_{>0}$ (Cohen and Peng, 2015). Iteratively applying m gives an algorithm that

2. We assume for simplicity that $\{a_i\}_{i\in[m]}$ is centrally symmetric, that is, $\{a_i\}_{i\in[m]} = -\{a_i\}_{i\in[m]}$.

computes an ε-estimate of σ_p, i.e., $\hat{w} \in \mathbb{R}^m_{>0}$ with $(1-\varepsilon)\sigma_p(A) \leq \hat{w} \leq (1+\varepsilon)\sigma_p(A)$ using $O(\frac{1}{1-|1-p/2|} \cdot \log(\log(m/\varepsilon))$ leverage score computations (Cohen and Peng, 2015).

However, efficiently obtaining ε-estimates of $\sigma_p(A)$ for $p \geq 4$ has been more challenging. Cohen and Peng (2015) showed that by applying the ellipsoid method to (4) estimates can be computed in $O(m \cdot \text{poly}(n) \log(1/\varepsilon))$ *time*. They also provided a recursive algorithm to compute high-accuracy estimates using $\Omega(n)$ leverage score computations. Additionally, Lee and Sidford (2019) shows how to compute ε-estimates using $O(\sqrt{n} \cdot p^2 \,\text{polylog}(mn/\varepsilon))$ leverage score computations. Their approach applies a descent method to a volumetric potential (equivalent to $\mathcal{F}_{\text{vec}}$ defined later up to a change of coordinates) that captures ℓ_p-Lewis weights. They show that the Hessian is stable around the minimizer which makes the convex objective function *locally* well conditioned. This ensures a $\log(1/\varepsilon)$-dependence of the descent method once weights are found that are close enough to the minimizer. To find such initial weights they used a homotopy method that slowly increases p.

Only recently, Fazel et al. (2022) provided the only known algorithms which compute ε-estimates of Lewis weights for $p > 2$ using a *nearly dimension-free* number of leverage score computations. Their method used $O(p^3 \log(mp/\varepsilon))$ leverage score computations. The derivation and analysis of their algorithms leveraged the following convex optimization problem, where we let $V := \text{diag}(v)$,

$$\min_{v \in \mathbb{R}^m_{>0}} \mathcal{F}_{\text{vec}}(v) \text{ where } \mathcal{F}_{\text{vec}}(v) := -\log\det(A^\top V A) + \frac{1}{1+\alpha_p} \mathbf{1}^\top v^{1+\alpha_p} \text{ for } \alpha_p := \frac{2}{p-2}. \quad (5)$$

Optimality conditions imply that its minimizer v_* satisfies $[v_*]_i^{\alpha_p} = a_i^\top (A^\top V A)^{-1} a_i$, and therefore $v_*^{1+\alpha_p} = \sigma_p(A)$.[3] Fazel et al. (2022) departed from contractivity analysis and instead performed an innovative, seemingly bespoke, analysis of (5).

The key insight of Fazel et al. (2022) is that a type of quasi-Newton step significantly decreases $\mathcal{F}_{\text{vec}}(v)$ when a geometrically motivated invariant holds. The invariant is $\rho_{\max}(v) \leq 1 + \alpha_p$, where

$$\rho_{\max}(v) := \max_{i \in [m]} \rho_i(v) \;\text{ where }\; \rho_i(v) := \frac{a_i^\top (A^\top V A)^{-1} a_i}{v_i^{\alpha_p}} = \frac{\sigma_i(V^{\frac{1}{2}} A)}{v_i^{1+\alpha_p}} \;\text{ for all }\; i \in [m]. \quad (6)$$

Note that $\rho_i(v_*) = 1$ for all $i \in [m]$, which means that the distance from $\rho(v)$ to the all-ones vector is a proxy for closeness to (rescaled) Lewis weights. The quantity $\rho_{\max}(v)$ has a geometric interpretation: $\{x \in \mathbb{R}^n : x^\top A^\top V A x \leq 1\} \subseteq \{x \in \mathbb{R}^n : \|V^{-\alpha_p/2} A x\|_\infty \leq \sqrt{\rho_{\max}(v)}\}$, which can be viewed as a notion of rounding (Fazel et al., 2022).[4]

To ensure that the geometric invariant is maintained, the authors introduced a rounding procedure. Fazel et al. (2022) provided an algorithm that uses $O(p^3 \log(mp/\varepsilon))$ leverage score computations and alternates between applying the rounding procedure and applying the quasi-Newton step. The quasi-Newton step can be written as updating v to v^+ where $v_i^+ = \left(1 + \eta \frac{\rho_i(v)-1}{\rho_i(v)+1}\right) v_i$, and the step-size η is $1/3$ for $p \geq 4$. Additionally, they provided another algorithm which avoids the rounding procedure by varying the step-size η per coordinate i depending on whether $\rho_i(v) \geq 1$ or $\rho_i(v) < 1$; it also uses $O(p^3 \log(mp/\varepsilon)$ leverage score computations.

There are additional algorithms that compute weaker approximations than ε-estimates of ℓ_p-Lewis weights (Cohen and Peng, 2015; Lee, 2016). To motivate these notions, recall that if w is

3. This rescaling of the coordinates of the Lewis weights to the $\frac{1}{1+\alpha_p} = 1 - \frac{2}{p}$ power is often convenient to work with and we use v rather than w to indicate vectors in this rescaled space.

4. Indeed, $a_i^\top (A^\top V A)^{-1} a_i \leq \rho_{\max}(v) v_i^{-\alpha_p}$. Hence, the rescaled vectors $\{a_i / (v_i^{\alpha_p/2} \sqrt{\rho_{\max}(v)})\}_{i \in [m]}$ belong to the ellipsoid $\{x \in \mathbb{R}^n : x^\top (A^\top V A)^{-1} x \leq 1\}$. The statement follows by considering the polar of each set.

an ℓ_p-Lewis weight vector, then it satisfies the fixed-point equation $w = \sigma(W^{\frac{1}{2}-\frac{1}{p}}A)$, and therefore $\|w\|_1 = \|\sigma(W^{\frac{1}{2}-\frac{1}{p}}A)\|_1 = n$. By relaxing the fixed-point equation to a one- or two-sided inequality, we arrive at the following (increasingly strong up to constants depending on p) notions of approximate ℓ_p-Lewis weights.

Definition 2 (Lewis weight approximations) *Let $A \in \mathbb{R}^{m\times n}$ be a non-degenerate, $w \in \mathbb{R}^m_{>0}$, $0 < \varepsilon < 1$, and $p > 0$. Then we say*

- *w is a* one-sided ε-approximation *of $\sigma_p(A)$ if $\sigma(W^{\frac{1}{2}-\frac{1}{p}}A) \le (1+\varepsilon)w$ and $\|w\|_1 \le (1+\varepsilon)n$.*
- *w is a* two-sided ε-approximation *of $\sigma_p(A)$ if $(1-\varepsilon)\sigma(W^{\frac{1}{2}-\frac{1}{p}}A) \le w \le (1+\varepsilon)\sigma(W^{\frac{1}{2}-\frac{1}{p}}A)$.*
- *w is an ε-*estimate *of $\sigma_p(A)$ if $(1-\varepsilon)\sigma_p(A) \le w \le (1+\varepsilon)\sigma_p(A)$.*

For many ℓ_p-embedding and -regression problems, the weakest one-sided approximation suffices, even when $\|w\|_1 = O(d)$, see Talagrand (1990); Cohen and Peng (2015); Woodruff and Yasuda (2023). Lee (2016) showed that iteratively applying the map $m(w)$ for $T = O(\log(m/n)/\varepsilon)$ iterations and outputting the average of the iterates results in a one-sided ε-approximation (Lee, 2016, Theorem 5.3.4). For some applications in optimization, however, a stronger notion of estimates are used (Lee and Sidford, 2019; Apers and Gribling, 2026). A natural question is how the various notions are related to each other. The only previously known conversion is that a two-sided ε-approximation is also an $O(\varepsilon p^2\sqrt{n})$-estimate (Fazel et al., 2022, Lemma 14).

1.2. Our results

In this paper we develop two new algorithms for computing ε-estimates of ℓ_p-Lewis weights for $p > 2$. Our algorithms use only $O(p^2\log(m/\varepsilon))$ leverage score computations, improving upon the prior nearly-dimension free results by a factor of p. (See Table 1.)

# Computes	Optimality	Reference
$O(p^3\log(m/\varepsilon))$	ε-estimate	Fazel et al. (2022)
$O(\log(m/n)/\varepsilon)$	ε-one-sided	Lee (2016)
$O(p^2\log(m/\varepsilon))$	ε-estimate	Algorithm 1
$O(p^2\log(m/\varepsilon))$	ε-estimate	Algorithm 2

Table 1: Comparison between the prior state of the art and our work, for the regime $p \ge 4$. The number of computes measures the number of leverage score computations.

Moreover, we show how to obtain these results by a fairly straightforward algorithm (the complete pseudocode is given later in Algorithm 1): starting from the all-ones vector $v^{(0)} = \mathbf{1}$, it performs the following iteration $T = O(p^2\log(m/\varepsilon))$ many times

$$v_i^{(t+1)} = \left(1 + \frac{\rho_i(v^{(t)})^{1/\alpha_p} - 1}{L}\right)v_i^{(t)}, \qquad \forall i \in [m] \tag{7}$$

where L is a suitably chosen step-size, and outputs $\widehat{w} = \hat{v}^{1+\alpha_p}$ for $\hat{v} = (a_i^\top(A^\top V^{(T)}A)^{-1}a_i)^{1/\alpha_p}$.

Theorem 3 *For $p > 2$, Algorithm 1 outputs an ε-estimate of $\sigma_p(A)$ in $O(p^2 \log(mp\alpha_p/\varepsilon))$ iterations. Each iteration computes the leverage scores of DA of some diagonal matrix D.*

Together with Cohen and Peng (2015), Theorem 3 gives the state-of-the-art rates for computing the ℓ_p-Lewis weights for all regimes of p.

Excitingly, rather than a particularly tailored analysis of a potential function, we analyze this algorithm using *relative smoothness and relative strong-convexity* (Lu et al., 2018), which are general regularity assumptions used in analyzing gradient-based methods for convex optimization, see also Bauschke et al. (2017); Tseng (2008). Via a simple extension, we prove that relatively smooth gradient descent converges at rates similar to those established in Lu et al. (2018) even when only a local variant of relative smoothness holds. We show how (7) is essentially equivalent to applying this method to a suitable objective. Additionally, we show that the convergence guarantees of this method directly correspond to computing ε-estimates of Lewis weights.

Complementing this result, we show that relative smoothness can also be applied directly to (5), the optimization problem considered in Fazel et al. (2022). We show that replacing (7) with

$$\boxed{v_i^{(t+1)} \leftarrow \left(1 + \frac{\rho_i(v^{(t)}) - 1}{L}\right)^{1/\alpha_p} v_i^{(t)}, \quad \forall i \in [m]} \tag{8}$$

for suitably chosen L optimizes $\mathcal{F}_{\text{vec}}(v)$ to accuracy $\varepsilon > 0$ in $O(p^2 \log(mp^2\alpha_p/\varepsilon))$ iterations. However, as in Fazel et al. (2022), significant work is needed to convert the convergence in function value to a guarantee on closeness to Lewis weights. We later provide Algorithm 2 which does this and analyze it in several steps. First, we show that the iterates from (8) in fact converge to one-sided approximations.

Theorem 4 *For $p > 2$, Algorithm 2 with parameter $\hat{\varepsilon}$ produces, after $T = O(p^2 \log(mp^2\alpha_p/\hat{\varepsilon}))$ iterations, a vector $w := [v^{(T)}]^{1+\alpha_p}$ that is a one-sided $\hat{\varepsilon}$-approximation of $\sigma_p(A)$.*

Then, we establish two new results that show how to convert a one-sided approximation to either a two-sided approximation or a multiplicative estimate, where $\bar{\beta}_p := \max\{1, 1/\alpha_p\} = \max\{1, \frac{p-2}{2}\}$.

Theorem 5 *For $p \geq 2$, if w is a one-sided ε_{one}-approximation of $\sigma_p(A)$ and $\widehat{w} := \sigma(w^{\frac{1}{2}-\frac{1}{p}})^{\frac{p}{2}}/w^{\beta_p}$, then $\widehat{w}$ is a two-sided ε_{two}-approximation of $\sigma_p(A)$ for $\varepsilon_{\text{two}} = 3\bar{\beta}_p n\varepsilon_{\text{one}}(1+\varepsilon_{\text{one}})^{\bar{\beta}_p}$.*

Theorem 6 *For $p > 2$, suppose w is a one-sided ε_{one}-approximation of $\sigma_p(A)$ satisfying*

$$\varepsilon_{\text{one}} \leq \frac{1}{\bar{\beta}_p n} \min\left\{\frac{1}{96(p-2)^2(4p-7)^2}, \frac{1}{50}\right\}.$$

Define $\widehat{w} \in \mathbb{R}^m_{>0}$ by $\widehat{w}_i := \sigma_i(w^{\frac{1}{2}-\frac{1}{p}})^{\frac{p}{2}}/w_i^{\beta_p}$ for each $i \in [m]$. Then $\widehat{w}$ is an ε_{est}-estimate of $\sigma_p(A)$, where $\varepsilon_{\text{est}} = 2(p-2)(4p-7)\sqrt{6\bar{\beta}_p n\varepsilon_{\text{one}}}$.

Applying Theorem 6 to the final iterate of Algorithm 2 yields ε-estimate as reflected in the following Theorem 7.

Theorem 7 *For $p > 2$, Algorithm 2 outputs an ε-estimate of $\sigma_p(A)$ in $O(p^2 \log(mp^2\alpha_p/\varepsilon))$ iterations. Each iteration computes the leverage scores of DA of some diagonal matrix D.*

Additionally, we establish two new results using our conversion tools. First, in Theorem 8 we give a postprocessing step that transforms any approximate minimizer v of $\mathcal{F}_{\text{vec}}$ satisfying $\rho_{\max}(v) \leq 1+\varepsilon$ into a two-sided approximation. Compared to the postprocessing step in Lemma 1 of Fazel et al. (2022), our approach does not incur a dimension-dependent polynomial factor loss in accuracy. Second, in Appendix E.4 we provide an improved analysis of a variant of (Lee, 2016, Algorithm 6), obtaining two-sided ε-approximations from $O(pn \log m/\varepsilon)$ approximate leverage-score computations to accuracy $O(\varepsilon/(pn))$.

Theorem 8 *For $p > 2$ and $\varepsilon \leq \min\{\frac{1}{1000}, \frac{1}{50\bar{\alpha}_p}\}$, suppose $v \in \mathbb{R}^m_{>0}$ satisfies $\mathcal{F}_{\text{vec}}(v) - \mathcal{F}_{\text{vec}}(v_*) \leq \varepsilon^3$ and $\rho_{\max}(v) \leq 1+\varepsilon$. Define $\widetilde{w} \in \mathbb{R}^m_{>0}$ coordinatewise by setting $\widetilde{w}_i = (\sigma_i(v)/v_i)^{1+1/\alpha_p}$ if $\rho_i(v) \leq 1-\varepsilon$, and $\widetilde{w}_i = v_i^{1+\alpha_p}$ otherwise. Then $\widetilde{w}$ is a two-sided $50\max\{\alpha_p, 1\}\varepsilon$-approximation of $\sigma_p(A)$.*

Though not the main focus of our work, we briefly discuss the runtime of our algorithms due to leverage score computations. Exact leverage scores of DA can be computed by first computing $G = A^\top D^2 A$ in time $O(mn^{\omega-1})$, then computing $H = G^{-1}A^\top D$ in time $O(mn^{\omega-1})$, and then computing the inner product of column i of H with row i of DA in time $O(mn)$ for all i. To the best of our knowledge, there is no better runtime to compute the leverage scores to high precision, though faster randomized algorithms for approximately computing leverage scores are known (Spielman and Srivastava, 2008; Clarkson and Woodruff, 2017). The conditioning of D affects this procedure through the required bit precision. Hence, we view controlling the range of D as an interesting open problem. We note that in both our algorithms, the diagonal scaling $D^{(t)} = (V^{(t)})^{1/2}$ changes by only a constant multiplicative factor in each coordinate between consecutive iterations, see Remarks 19, 33.

1.3. Approach

Here we provide a brief overview of our approach. First, we briefly sketch the relative smoothness and convexity framework. (Formal definitions are deferred to Section 2.) For differentiable functions f and h, we say that f is μ-strongly convex and L-smooth relative to h when

$$\mu D_h(x,y) \leq f(x) - f(y) - \langle \nabla f(y), x-y\rangle \leq L D_h(x,y) \qquad \forall x, y, \tag{9}$$

where $D_h(x,y) := h(x) - h(y) - \langle \nabla h(y), x-y\rangle$ is the Bregman divergence associated with h. For twice-differentiable f and h (9) is equivalent to $\mu\nabla^2 h(x) \preceq \nabla^2 f(x) \preceq L\nabla^2 h(x)$ for all x (Lu et al., 2018). Lu et al. (2018), roughly, shows that, when $0 < \mu < L$, the gradient descent scheme

$$x^{(t+1)} \leftarrow \operatorname*{argmin}_{x\in\mathcal{C}} \left\{ f(x^{(t)}) + \langle \nabla f(x^{(t)}), x - x^{(t)}\rangle + L D_h(x, x^{(t)}) \right\}$$

converges linearly at a rate $1 - \frac{\mu}{L}$ in function value, and in Bregman distance to the minimizer, e.g., $D_h(x^*, x^{(t)}) \leq \frac{L}{\mu}\left(1 - \frac{\mu}{L}\right)^t D_h(x^*, x^{(0)})$ where x^* is the minimizer of f.

In a nutshell, the update step in (7) arises from a careful extension of this framework applied to the convex optimization problem over positive definite n-by-n matrices

$$\begin{aligned}&\min_{M\in\mathcal{S}^n_{\succ 0}} \mathcal{F}_{\mathrm{mat}}(M), \text{ where}\\ &\mathcal{F}_{\mathrm{mat}}(M) := -\operatorname{logdet}(M) + \frac{1}{1+\beta_p}\sum_{i\in[m]}(a_i^\top M a_i)^{1+\beta_p} \text{ for } \beta_p := \frac{p-2}{2}\,.\end{aligned} \tag{10}$$

This problem is essentially dual to (5); see Lemma 12 and note that $\beta_p = \alpha_p^{-1}$. The minimizer M_* of (10) satisfies $M_*^{-1} = A^\top \overline{V} A$, where $\overline{V} := \sigma_p(A)^{\frac{1}{1+\alpha_p}}$, and thus encodes the Lewis weights. The update from (7) corresponds to the gradient descent method applied the above objective where $M^{-1} = A^\top V A$. It is easy to see that $\mathcal{F}_{\mathrm{mat}}$ is 1-strongly convex relative to $h_{\mathrm{mat}}(M) = -\operatorname{logdet}(M)$. Moreover, the Bregman divergence associated to h_{mat} roughly measures spectral closeness between matrices: $D_{h_{\mathrm{mat}}}(M_*, M) \le \varepsilon^2/4$ implies $(1-\varepsilon)M_* \preceq M \preceq (1+\varepsilon)M_*$ (Lemma 20), which shows that near-optimal points of $\mathcal{F}_{\mathrm{mat}}$ provide ε-estimates of the Lewis weights.

The only remaining challenge is to establish the relative smoothness of $\mathcal{F}_{\mathrm{mat}}$ with respect to h_{mat}. Unfortunately, a sufficient global bound is unknown, even for sub-level sets. Instead, we introduce *local relative smoothness between iterates*, a straightforward extension of relative smoothness that just holds between the iterates. The idea of using different local (or adaptive) notions of smoothness has been used before, e.g., in Sidford and Tian (2018); Malitsky and Mishchenko (2020); Latafat et al. (2025). In particular, Li et al. (2018) developed a ball-local version of relative smoothness, requiring the relative-smoothness inequality to hold uniformly within a neighborhood of each point, while Godeme et al. (2023) used local relative strong convexity on a prescribed neighborhood, typically around a solution.

We prove in Section 2 that such local relative smoothness suffices for linear convergence. In particular, we show that when $M = (A^\top V A)^{-1}$ for some $V = \operatorname{Diag}(v)$ with $v \in \mathbb{R}^m_{>0}$, then

$$\nabla^2 \mathcal{F}_{\mathrm{mat}}(M) \preceq (1 + \beta_p \Phi_{\max}(v))\, \nabla^2 h_{\mathrm{mat}}(M).$$

where

$$\Phi_{\max}(v) := \max_{i\in[m]} \Phi_i(v) \text{ where } \Phi_i(v) := \frac{(a_i^\top (A^\top V A)^{-1} a_i)^{\beta_p}}{v_i} = \rho_i(v)^{1/\alpha_p} \text{ for all } i \in [m]\,. \tag{11}$$

Applying this gradient descent scheme to $\mathcal{F}_{\mathrm{mat}}$ and h_{mat} results in the iterates (7), when written in terms of v. To establish convergence, we set $L = 32p\max\{\beta_p, 1\}$ and show that $\Phi_{\max}(v)$ is uniformly bounded in the segment between each pair of iterates $v^{(t)}$ and $v^{(t+1)}$.

As discussed, to further showcase the approach, we then apply the same local relative smoothness framework to the potential $\mathcal{F}_{\mathrm{vec}}$ used in Lee and Sidford (2019); Fazel et al. (2022). In this case, it is easy to see that $\mathcal{F}_{\mathrm{vec}}$ is 1-strongly convex relative to $h_{\mathrm{vec}}(v) = \frac{1}{1+\alpha_p}\mathbf{1}^\top v^{1+\alpha_p}$. We show that for any $v \in \mathbb{R}^m_{>0}$, we have

$$\nabla^2 \mathcal{F}_{\mathrm{vec}}(v) \preceq \left(1 + \alpha_p^{-1}\rho_{\max}(v)\right) \nabla^2 h_{\mathrm{vec}}(v).$$

In a similar fashion as for (7), we establish *local* relative smoothness between iterates when $L = 32p/\alpha_p$, thus establishing convergence in function value. As discussed earlier, with more work

we are able to use this to obtain approximations of Lewis weights. One component of this reduction is an efficient conversion of one-sided approximations into two-sided approximations (see Appendix E). We do so by defining a transformation $w \mapsto \widehat{w}$ such that $\|\rho(\widehat{w}^{\frac{1}{2}-\frac{1}{p}}) - 1\|_\infty$ is controlled by the spectral approximation quality of $A^\top W^{\frac{1}{2}-\frac{1}{p}} A$ by $A^\top \widehat{W}^{\frac{1}{2}-\frac{1}{p}} A$, which can in turn be bounded by $\|\widehat{w} - \sigma(w^{\frac{1}{2}-\frac{1}{p}})\|_1$ and is small whenever w is a one-sided approximation.

1.4. Paper organization

In the remainder of this paper we extend the relative smoothness framework (Section 2), apply it to $\mathcal{F}_{\text{mat}}$ (Section 3) and $\mathcal{F}_{\text{vec}}$ (Appendix D), and establish conversion results between the notions of approximation (Appendix E). We conclude this introduction by providing our notation.

General notation. We use lowercase letters for vectors and capital letters for matrices. When the context is clear, a capital letter additionally denotes the diagonal matrix formed from its lowercase counterpart, e.g., $W = \text{Diag}(w)$. The all-zero and all-one vectors of appropriate dimension are denoted by $\mathbf{0}, \mathbf{1}$, respectively. For $u, v \in \mathbb{R}^n$, we write $u \leq v$ to denote the entrywise inequality $u_i \leq v_i$ for all $i \in [n]$, and write $u \approx_\varepsilon v$ if $(1-\varepsilon)u_i \leq v_i \leq (1+\varepsilon)u_i$ for all $i \in [n]$. We use $\text{Diag}(u)$ to denote the diagonal matrix with entries $\text{Diag}(u)_{ii} = u_i$. For any matrices A, B, we write $A \succeq B$, or equivalently, $B \preceq A$ when $A - B$ is positive semidefinite. Moreover, we define $\langle A, B\rangle \coloneqq \text{tr}(A^\top B)$. We write $\mathcal{S}^n$ for the space of symmetric n-by-n matrices. For any convex set $\mathcal{C}$, we use $\text{int}\,\mathcal{C}$ to denote its interior. We use $\otimes$ to denote the Kronecker product.

Lewis weight notation. For a matrix $A \in \mathbb{R}^{m\times n}$, we write $v_*(A) \coloneqq \sigma_p(A)^{\frac{1}{1+\alpha_p}}$ when p is clear from context, and write v_* when the underlying matrix is clear from the context. Denote $V_* = \text{diag}(v_*)$. For any $p > 2$, we denote $\alpha_p = \frac{2}{p-2}$, $\beta_p = 1/\alpha_p$, $\bar{\alpha}_p = \max\{1, \alpha_p\}$, and $\bar{\beta}_p = \max\{1, \beta_p\}$.

2. Locally relatively smooth gradient descent framework

In this section, we present a straightforward local extension of the relative smoothness framework introduced in Lu et al. (2018), where the relative smoothness condition only holds locally, along linear combinations of selected pairs of points. This differs from the local relative smoothness framework of Li et al. (2018), in which relative smoothness holds when restricted to a ball region around any given point. The proofs of the claims in the section are deferred to Appendix B.

Definition 9 (Local relative smoothness) *Let $f, h : \mathcal{C} \to \mathbb{R}$ be differentiable functions on a convex set $\mathcal{C}$, and let $x, y \in \text{int}\,\mathcal{C}$. We say that f is L-smooth relative to h between x, y if we have*

$$f((1-\lambda)x + \lambda y) \leq f(x) + (1-\lambda)\langle \nabla f(x), y - x\rangle + LD_h((1-\lambda)x + \lambda y, x), \quad \forall \lambda \in [0,1].$$

If f and h are twice differentiable, this condition is equivalent to

$$\nabla^2 f((1-\lambda)x + \lambda y) \preceq L\nabla^2 h((1-\lambda)x + \lambda y), \quad \forall \lambda \in [0,1].$$

We show that for any objective function f defined on $\mathcal{C}$ that is μ-strongly convex relative to some known convex function h, repeatedly performing the following update

$$x^{(t+1)} \leftarrow \underset{x\in\mathcal{C}}{\text{argmin}} \left\{ f(x^{(t)}) + \langle \nabla f(x^{(t)}), x - x^{(t)}\rangle + LD_h(x, x^{(t)}) \right\} \tag{12}$$

converges to the minimizer of f, given that f is L-smooth relative to h between each $x^{(t)}$ and $x^{(t+1)}$.

Proposition 10 *Let $f, h : \mathcal{C} \to \mathbb{R}$ be differentiable functions on a convex set $\mathcal{C}$ where f is μ-strongly convex relative to h for some $\mu \geq 0$, and h is convex. If in the updating scheme (12) there exists $L > 0$ such that f is L-smooth relative to h between $x^{(t)}$ and $x^{(t+1)}$ for every iteration t then*

$$D_h(x, x^{(t)}) + \frac{1}{L}\sum_{k\in[t]}\left(1-\frac{\mu}{L}\right)^{t-k}(f(x^{(k)}) - f(x)) \leq \left(1-\frac{\mu}{L}\right)^{t} D_h(x, x^{(0)}), \quad \forall x \in \mathcal{C}, t \in \mathbb{N}^*. \tag{13}$$

Consequently, for $x^* \coloneqq \operatorname{argmin}_{x\in\mathcal{C}} f(x)$ and when $\mu > 0$,

$$D_h(x^*, x^{(t)}) \leq \left(1-\frac{\mu}{L}\right)^{t} D_h(x^*, x^{(0)}) \text{ and } f(x^{(t)}) - f(x^*) \leq \frac{\mu(1-\mu/L)^t}{1-(1-\mu/L)^t}\cdot D_h(x^*, x^{(0)}).$$

The proof of Proposition 10 is inspired by Theorem 3.1 of Lu et al. (2018). Our main contribution is to extend their analysis to the setting in which relative smoothness holds only locally rather than globally, and to measure convergence using both the Bregman distance to the minimizer as well as the function value gap. The key step for proving Proposition 10 is to establish the following lemma.

Lemma 11 *In the setting of Proposition 10, for each iteration t and any $x \in \mathcal{C}$ we have*

$$D_h(x, x^{(t+1)}) \leq \left(1-\frac{\mu}{L}\right) D_h(x, x^{(t)}) + \frac{1}{L}(f(x) - f(x^{(t+1)})).$$

3. A relative smoothness algorithm based on a matrix potential $\mathcal{F}_{\text{mat}}$

In this section, we present an algorithm that computes ε-estimates of Lewis weights by approximately solving (10) via the locally relatively smooth gradient descent framework in Section 2. Throughout the section, we let $h_{\text{mat}}(M) \coloneqq -\operatorname{logdet}(M)$ for all $M \in \mathcal{S}^n_{\succ 0}$, and denote $\bar{\varepsilon} = \frac{\varepsilon}{2(1+\alpha_p)}$.

3.1. Properties of $\mathcal{F}_{\text{mat}}$

Here we present several properties of $\mathcal{F}_{\text{mat}}$, including its duality with $\mathcal{F}_{\text{vec}}$, explicit formulas for its gradient, Hessian and optimum, and its relative strong convexity and local smoothness properties with respect to h_{mat}. The proof of Lemma 12 and Lemma 13 can be found in Appendix C.

Lemma 12 *The optimization problems (5) and (10) are dual to each other in the following sense:*

$$\min_{M\succ 0} \mathcal{F}_{\text{mat}}(M) = n - \min_{w>0} \mathcal{F}_{\text{vec}}(w)$$

Lemma 13 *For any $M \in \mathcal{S}^n_{\succ 0}$ the gradient and Hessian of $\mathcal{F}_{\text{mat}}$ have the following expressions:*

$$\nabla \mathcal{F}_{\text{mat}}(M) = -M^{-1} + \sum_{i\in[m]} (a_i^\top M a_i)^{\beta_p} a_i a_i^\top \text{ and} \tag{14}$$

$$\nabla^2 \mathcal{F}_{\text{mat}}(M) = M^{-1}\otimes M^{-1} + \beta_p \sum_{i\in[m]} (a_i^\top M a_i)^{\beta_p - 1} a_i a_i^\top \otimes a_i a_i^\top .$$

Moreover, $\mathcal{F}_{\text{mat}}$ has a unique minimizer M_* in $\mathcal{S}^n_{\succ 0}$ that satisfies $M_*^{-1} = A^\top V_* A$.

Lemma 14 *$\mathcal{F}_{\mathrm{mat}}$ is 1-relatively strongly convex with respect to h_{mat}. Moreover, if $M = (A^\top V A)^{-1}$ for some $V = \mathrm{Diag}(v)$ with $v \in \mathbb{R}^m_{>0}$, we have*

$$\nabla^2 \mathcal{F}_{\mathrm{mat}}(M) \preceq (1 + \beta_p \Phi_{\max}(v))\, \nabla^2 h_{\mathrm{mat}}(M). \tag{15}$$

Proof The 1-relative strong convexity of $\mathcal{F}$ with respect to h_{mat} follows from the convexity of $(a_i^\top M a_i)^{1+\beta_p} = (a_i^\top M a_i)^{p/2}$. More formally, observe that

$$\nabla^2 \mathcal{F}_{\mathrm{mat}}(M) = \nabla^2 h_{\mathrm{mat}}(M) + \frac{1}{\alpha_p} \sum_{i \in [m]} (a_i^\top M a_i)^{\beta_p - 1} a_i a_i^\top \otimes a_i a_i^\top \succeq \nabla^2 h_{\mathrm{mat}}(M),$$

proving that $\mathcal{F}_{\mathrm{mat}}$ is 1-relatively strongly convex with respect to $h_{\mathrm{mat}}(M)$.

We now establish (15). First, using $M^{1/2} a_i a_i^\top M^{1/2} \preceq a_i^\top M a_i I$ and monotonicity of the Kronecker product, we observe that

$$\begin{aligned}
&\sum_{i \in [m]} (a_i^\top M a_i)^{\beta_p - 1} M^{1/2} a_i a_i^\top M^{1/2} \otimes M^{1/2} a_i a_i^\top M^{1/2} \\
&\qquad \preceq M^{1/2} \left(\sum_{i \in [m]} (a_i^\top M a_i)^{\beta_p} a_i a_i^\top \right) M^{1/2} \otimes I.
\end{aligned}$$

If $M = (A^\top V A)^{-1}$ for some $V = \mathrm{Diag}(v)$, we have

$$(a_i^\top M a_i)^{\beta_p} = (a_i^\top (A^\top V A)^{-1} a_i)^{\beta_p} = \Phi_i(v) v_i \tag{16}$$

and therefore

$$\begin{aligned}
&\sum_{i \in [m]} (a_i^\top M a_i)^{\beta_p - 1} M^{1/2} a_i a_i^\top M^{1/2} \otimes M^{1/2} a_i a_i^\top M^{1/2} \\
&\qquad \preceq M^{1/2} \left(\sum_{i \in [m]} v_i \Phi_i(v) a_i a_i^\top \right) M^{1/2} \otimes I \\
&\qquad \preceq \Phi_{\max}(v) M^{1/2} \left(\sum_{i \in [m]} v_i a_i a_i^\top \right) M^{1/2} \otimes I \\
&\qquad = \Phi_{\max}(v) I \otimes I\,.
\end{aligned}$$

(15) then follows as $\nabla^2(-\log\det(M)) = M^{-1} \otimes M^{-1}$ and therefore

$$\sum_{i \in [m]} (a_i^\top M a_i)^{\beta_p - 1} a_i a_i^\top \otimes a_i a_i^\top \preceq \Phi_{\max}(v) \nabla^2 h_{\mathrm{mat}}(M).$$

■

Lemma 15 *We have $D_{h_{\mathrm{mat}}}\big((A^\top V_* A)^{-1}, (A^\top A)^{-1}\big) \leq m - n$.*

Proof By the definition of the Bregman divergence, we have

$$D_{h_{\mathrm{mat}}}\left((A^\top V_* A)^{-1}, (A^\top A)^{-1}\right) \\ = -\operatorname{logdet}((A^\top V_* A)^{-1}) + \operatorname{logdet}((A^\top A)^{-1}) + \left\langle A^\top A, (A^\top V_* A)^{-1} - (A^\top A)^{-1}\right\rangle.$$

Using $A^\top V_* A \preceq A^\top A$, we have

$$-\operatorname{logdet}((A^\top V_* A)^{-1}) + \operatorname{logdet}((A^\top A)^{-1}) = \operatorname{logdet}(A^\top V_* A) - \operatorname{logdet}(A^\top A) \le 0.$$

Moreover, since $[v_*]_i = [\sigma_p(A)]_i^{\frac{\beta_p}{1+\beta_p}} \le 1$ for any $i \in [m]$, we have

$$\left\langle A^\top A, (A^\top V_* A)^{-1} - (A^\top A)^{-1}\right\rangle = \operatorname{tr}[A(A^\top V_* A)^{-1} A^\top] - n = \sum_{i\in[m]} [v_*]_i^{1/\beta_p} - n \le m - n.$$

Hence, we can conclude that $D_{h_{\mathrm{mat}}}\left((A^\top V_* A)^{-1}, (A^\top A)^{-1}\right) \le m - n$. ■

3.2. Applying the local relative smoothness framework to $\mathcal{F}_{\mathrm{mat}}$

Here we give our algorithm that solves (10) by iteratively performing the update in (7). Throughout, our iterate $M^{(t)}$ is of the form $M^{(t)} := (A^\top V^{(t)} A)^{-1}$, where $V^{(t)} = \operatorname{Diag}(v^{(t)})$ for $v^{(t)} \in \mathbb{R}^m_{>0}$. Lemma 16 shows that the update from (7) in terms of $v^{(t)}$ follows the gradient descent scheme.

Algorithm 1: High-precision algorithm using the matrix potential $\mathcal{F}_{\mathrm{mat}}$

Input: non-degenerate $A \in \mathbb{R}^{m\times n}$, $p > 2$, $\varepsilon > 0$

Set $L = 32p\bar{\beta}_p$, $T = \lceil 4L \log(2m/\bar{\varepsilon})\rceil$ where $\bar{\varepsilon} = \frac{\varepsilon}{2(1+\alpha_p)}$, and $v_i^{(0)} = 1$ for all $i \in [m]$.

for $t = 0, 1, \ldots, T-1$ **do**

  $v_i^{(t+1)} \leftarrow \left(1 + \frac{\Phi_i(v^{(t)})-1}{L}\right) v_i^{(t)}, \forall i \in [m]$ `// Recall` $\Phi_i(v) = \frac{(a_i^\top (A^\top V A)^{-1} a_i)^{\beta_p}}{v_i}$.

end

Return $\widehat{w} \in \mathbb{R}^m_{>0}$, where $\widehat{w}_i = (a_i^\top (A^\top V^{(T)} A)^{-1} a_i)^{1+\beta_p}$.

Theorem 3 *For $p > 2$, Algorithm 1 outputs an ε-estimate of $\sigma_p(A)$ in $O(p^2 \log(mp\alpha_p/\varepsilon))$ iterations. Each iteration computes the leverage scores of DA of some diagonal matrix D.*

Lemma 16 *For any iteration t in Algorithm 1, we have*[5]

$$M^{(t+1)} = \operatorname*{argmin}_{M\succeq 0} \left\{ \mathcal{F}_{\mathrm{mat}}(M^{(t)}) + \langle \nabla \mathcal{F}_{\mathrm{mat}}(M^{(t)}), M - M^{(t)}\rangle + L D_{h_{\mathrm{mat}}}(M, M^{(t)}) \right\}$$

Proof Given the choice of h, we have that

$$D_h(M, M^{(t)}) = -\operatorname{logdet}(M) + \operatorname{logdet}(M^{(t)}) + \langle [M^{(t)}]^{-1}, M - M^{(t)}\rangle \tag{17}$$

5. Since $\mathcal{F}_{\mathrm{mat}}$ is convex and $\lim_{M\to\partial\{M\mid M\succeq 0\}} \mathcal{F}_{\mathrm{mat}}(M) = +\infty$, the minimizer is always attained in the interior.

and, by Lemma 13,

$$\begin{aligned}
&\nabla \mathcal{F}_{\text{mat}}(M^{(t)}) + L\nabla D_h(M, M^{(t)})\big|_{M=M^{(t+1)}} \\
&\quad = \nabla \mathcal{F}_{\text{mat}}(M^{(t)}) + L\big(-[M^{(t+1)}]^{-1} + [M^{(t)}]^{-1}\big) \\
&\quad = (L-1)[M^{(t)}]^{-1} + \sum_{i\in[m]} (a_i^\top M^{(t)} a_i)^{\beta_p} a_i a_i^\top - L[M^{(t+1)}]^{-1}\,.
\end{aligned} \tag{18}$$

Substituting $M^{(t)} = (A^\top V^{(t)} A)^{-1}$ for any t, we have $(a_i^\top M^{(t)} a_i)^{\beta_p} = \Phi_i(v^{(t)}) v_i^{(t)}$ by (16). Then, Eq. (18) equals

$$\sum_{i\in[m]} \left((L + \Phi_i(v^{(t)}) - 1)v_i^{(t)} - L v_i^{(t+1)}\right) a_i a_i^\top = 0$$

Since $D_{h_{\text{mat}}}(M, M^{(t)})$ is convex, we can conclude that

$$M^{(t+1)} = \operatorname*{argmin}_{M\succeq 0} \left\{ \mathcal{F}_{\text{mat}}(M^{(t)}) + \langle \nabla \mathcal{F}_{\text{mat}}(M^{(t)}), M - M^{(t)}\rangle + L D_{h_{\text{mat}}}(M, M^{(t)}) \right\}.$$

■

Lemma 17 *For any iteration t in Algorithm 1, if $\Phi_{\max}(v^{(t)}) \le 4\bar{\beta}_p \le L$ then $\Phi_{\max}(v^{(t+1)}) \le 4\bar{\beta}_p$.*

Proof By the update formula in Line 3, we have $v_i^{(t+1)} \ge \left(1 - \frac{1}{L}\right) v_i^{(t)}$, which leads to

$$\begin{aligned}
\Phi_i(v^{(t+1)}) &= \frac{v_i^{(t)}}{v_i^{(t+1)}} \cdot \frac{(a_i^\top (A^\top V^{(t+1)} A)^{-1} a_i)^{\beta_p}}{v_i^{(t)}} \\
&\le \left(1 - \frac{1}{L}\right)^{-\beta_p} \frac{v_i^{(t)}}{v_i^{(t+1)}} \cdot \frac{(a_i^\top (A^\top V^{(t)} A)^{-1} a_i)^{\beta_p}}{v_i^{(t)}} \\
&= \left(1 - \frac{1}{L}\right)^{-\beta_p} \left(1 + \frac{\Phi_i(v^{(t)}) - 1}{L}\right)^{-1} \cdot \Phi_i(v^{(t)})\,.
\end{aligned}$$

Since $L > 1$, the function $\psi\colon \mathbb{R}^+ \to \mathbb{R}$ defined as $\psi(x) := \left(1 - \frac{1}{L}\right)^{-\beta_p} \left(1 + \frac{x-1}{L}\right)^{-1} x$, is monotonically increasing for $x > 0$. Then, using $\Phi_{\max}(v^{(t)}) \le 4\bar{\beta}_p$, we have

$$\Phi_{\max}(v^{(t+1)}) \le \left(1 - \frac{1}{L}\right)^{-\beta_p} \left(1 + \frac{4\bar{\beta}_p - 1}{L}\right)^{-1} 4\bar{\beta}_p \le \left(1 - \frac{\beta_p}{L}\right)^{-1} \left(1 + \frac{4\bar{\beta}_p}{2L}\right)^{-1} 4\bar{\beta}_p \le 4\bar{\beta}_p,$$

where the second inequality uses $L \ge 4\bar{\beta}_p$ and $4\bar{\beta}_p \ge 2$, and the fact that $(1-x)^{-\beta_p} \le (1-\beta_p x)^{-1}$ for all $0 \le x \le 1/(2\beta_p)$. The third inequality uses $\frac{\beta_p}{L} \le \frac{1}{4}$, and the fact that $(1-x)^{-1}(1+2x)^{-1} \le 1$ for all $0 \le x \le 1/4$. ■

Lemma 18 *For any iteration t in Algorithm 1 and any $\lambda \in [0,1]$, if $\Phi_{\max}(v^{(t)}) \le 4\bar{\beta}_p$ and $L \ge 16p\bar{\beta}_p$, we have that*

$$\nabla^2 \mathcal{F}_{\text{mat}}(M_\lambda) \preceq \left(1 + 16p\bar{\beta}_p\right) \nabla^2 h_{\text{mat}}(M_\lambda) \quad \textit{where} \quad M_\lambda := (1-\lambda)M^{(t)} + \lambda M^{(t+1)}\,. \tag{19}$$

Proof By the update formula in Line 3, we have $\frac{|v_i^{(t+1)}-v_i^{(t)}|}{v_i^{(t)}} \le \frac{|\Phi_i(v^{(t)})-1|}{L} \le \frac{4\beta_p-1}{L} \le \frac{1}{4}$ where the second inequality follows from $\Phi_{\max}(v^{(t)}) \le 4\bar{\beta}_p$, and the last inequality uses $L \ge 16p\bar{\beta}_p \ge 16\bar{\beta}_p$. Consequently, $\frac{3}{4}A^\top V^{(t)}A \preceq A^\top V^{(t+1)}A \preceq \frac{5}{4}A^\top V^{(t)}A$, and thus $\frac{4}{5}M^{(t)} \preceq M_\lambda \preceq \frac{4}{3}M^{(t)}$ for any $\lambda \in [0,1]$. By Lemma 13, the Hessian of $\mathcal{F}_{\mathrm{mat}}$ admits the decomposition

$$\nabla^2 \mathcal{F}_{\mathrm{mat}}(M_\lambda) = M_\lambda^{-1} \otimes M_\lambda^{-1} + \frac{1}{\alpha_p}\sum_{i\in[m]} (a_i^\top M_\lambda a_i)^{\beta_p - 1} a_i a_i^\top \otimes a_i a_i^\top.$$

We bound the second term as follows:

$$\begin{aligned}
&\sum_{i\in[m]} (a_i^\top M_\lambda a_i)^{\beta_p-1} M_\lambda^{1/2} a_i a_i^\top M_\lambda^{1/2} \otimes M_\lambda^{1/2} a_i a_i^\top M_\lambda^{1/2} \\
&\quad \preceq 2\sum_{i\in[m]} (a_i^\top M_\lambda a_i)^{\beta_p-1} [M^{(t)}]^{1/2} a_i a_i^\top [M^{(t)}]^{1/2} \otimes [M^{(t)}]^{1/2} a_i a_i^\top [M^{(t)}]^{1/2} \\
&\quad \preceq 2[M^{(t)}]^{1/2} \left(\sum_{i\in[m]} (a_i^\top M_\lambda a_i)^{\beta_p} a_i a_i^\top\right) [M^{(t)}]^{1/2} \otimes I. \qquad (20)
\end{aligned}$$

where the first inequality uses the spectral closeness between M_λ and $M^{(t)}$. Next, by Lemma 17, for each coordinate i we have

$$\begin{aligned}
(a_i^\top M_\lambda a_i)^{\beta_p} &\le (\max\{a_i^\top M^{(t)} a_i, a_i^\top M^{(t+1)} a_i\})^{\beta_p} \\
&\le \max\left\{v_i^{(t)}\Phi_i(v^{(t)}), v_i^{(t+1)}\Phi_i(v^{(t+1)})\right\} \le 8\bar{\beta}_p v_i^{(t)}.
\end{aligned}$$

Therefore,

$$[M^{(t)}]^{1/2}\left(\sum_{i\in[m]} (a_i^\top M_\lambda a_i)^{\beta_p} a_i a_i^\top\right)[M^{(t)}]^{1/2} \preceq 8\bar{\beta}_p [M^{(t)}]^{1/2}\left(\sum_{i\in[m]} v_i^{(t)} a_i a_i^\top\right)[M^{(t)}]^{1/2} = 8\bar{\beta}_p I,$$

which combined with Eq. (20) gives $\sum_{i\in[m]} (a_i^\top M_\lambda a_i)^{\beta_p-1} M_\lambda^{1/2} a_i a_i^\top M_\lambda^{1/2} \otimes M_\lambda^{1/2} a_i a_i^\top M_\lambda^{1/2} \preceq 16\bar{\beta}_p I \otimes I$. Consequently, $\nabla^2 \mathcal{F}_{\mathrm{mat}}(M_\lambda) \preceq \left(1 + 16p\bar{\beta}_p\right) \nabla^2 h_{\mathrm{mat}}(M_\lambda)$. ■

Proof of Theorem 3. By Lemma 16, each iteration of Algorithm 1 can be equivalently written as

$$M^{(t+1)} = \operatorname*{argmin}_{M \succ 0} \left\{\mathcal{F}_{\mathrm{mat}}(M^{(t)}) + \langle \nabla \mathcal{F}_{\mathrm{mat}}(M^{(t)}), M - M^{(t)}\rangle + L D_h(M, M^{(t)})\right\},$$

where $M^{(t)} = (A^\top V^{(t)} A)^{-1}$. Since

$$\Phi_{\max}(v^{(0)}) = \left(\max_{i\in[m]} \frac{\sigma_i(v^{(0)})}{[v_i^{(0)}]^{1+1/\beta_p}}\right)^{\beta_p} = \max_{i\in[m]} \sigma_i(v^{(0)})^{\beta_p} \le 1,$$

Lemma 17 implies that $\Phi_{\max}(v^{(t)}) \le 4\bar{\beta}_p$ for all iterations t. By Lemma 18, it follows that $\mathcal{F}_{\mathrm{mat}}$ is $(1 + 16p\bar{\beta}_p) \le L$-smooth relative to h between any two consecutive iterates $M^{(t)}$ and $M^{(t+1)}$.

Moreover, since $\mathcal{F}_{\mathrm{mat}}$ is 1-strongly convex relative to h by Lemma 14, Proposition 10 yields

$$D_h(M_*, M^{(T)}) \le \left(1 - \frac{1}{L}\right)^T D_h(M_*, M^{(0)}) + \frac{1}{L} \sum_{t \in [T]} \left(1 - \frac{1}{L}\right)^{T-t} (\mathcal{F}_{\mathrm{mat}}(M_*) - \mathcal{F}_{\mathrm{mat}}(M^{(t)}))$$
$$\le \left(1 - \frac{1}{L}\right)^T D_h(M_*, M^{(0)}) \le \frac{\bar{\varepsilon}^2}{16}$$

where the second inequality uses that $M_* = \arg\min_{M \succ 0} \mathcal{F}_{\mathrm{mat}}(M)$. Lemma 20 then implies $M^{(T)} \approx_{\bar{\varepsilon}/2} M_*$. Consequently, $\widehat{v} \in \mathbb{R}^m_{>0}$ with $\widehat{v}_i = (a_i^\top (A^\top V^{(T)} A)^{-1} a_i)^{\beta_p}$ for all $i \in [m]$ satisfies

$$\left| \frac{\widehat{v}_i}{[v_*]_i} - 1 \right| = \left| \frac{a_i^\top M^{(T)} a_i}{a_i^\top M_* a_i} - 1 \right| \le \frac{\bar{\varepsilon}}{2} \text{ for all } i \in [m] .$$

Therefore $(1 - \varepsilon)\sigma_p(A) \le \widehat{w} \le (1 + \varepsilon)\sigma_p(A)$ since $\sigma_p(A) = v_*^{1+1/\beta_p}$ and $\widehat{w} = \widehat{v}^{1+1/\beta_p}$. ■

Remark 19 *By Lemma 17 and the initialization, $\Phi_{\max}(v^{(t)}) \le 4\bar{\beta}_p$ for every iteration t in Algorithm 1. Therefore, for every coordinate $i \in [m]$,*

$$\left| \frac{v_i^{(t+1)}}{v_i^{(t)}} - 1 \right| = \frac{|\Phi_i(v^{(t)}) - 1|}{L} \le \frac{4\bar{\beta}_p}{32p\bar{\beta}_p} = \frac{1}{8p} \le \frac{1}{4}.$$

Consequently, the diagonal scaling $D^{(t)} = (V^{(t)})^{1/2}$ used in each leverage score computation satisfies

$$\sqrt{\frac{3}{4}}\, D_{ii}^{(t)} \le D_{ii}^{(t+1)} \le \sqrt{\frac{5}{4}}\, D_{ii}^{(t)}, \quad \forall i \in [m].$$

4. Conclusion

In this paper we provide two algorithms for computing approximations of ℓ_p-Lewis weights. Additionally, we provide simple procedures that convert weaker notions of approximation into stronger ones, e.g., that turn one-sided approximations into two-sided approximations. For the fundamental problem of computing ε-estimates, our methods improve upon the prior state-of-the-art by a factor of p. Moreover, we obtain these algorithms by a general *locally* relatively smooth gradient descent method and straightforward applications of it to convex formulations of Lewis weights.

Altogether, these algorithms and the analysis shed light on the complexity of ℓ_p-Lewis weight computation, through the lens of relative smoothness and strong convexity. Given the fundamental and pervasive nature of ℓ_p-Lewis weights and how natural the associated objective functions are, we hope this work may facilitate the development of efficient optimization algorithms more broadly.

Acknowledgments

We thank Simon Apers for many useful discussions during the development of this work. We thank anonymous reviewers from COLT 2026 for their feedback and LLMs for writing advice. Aaron Sidford was supported in part by a Microsoft Research Faculty Fellowship, NSF CAREER Grant CCF1844855, NSF Grant CCF-1955039, and a PayPal research award. Chenyi Zhang was supported by a Shoucheng Zhang Graduate Fellowship.

References

Simon Apers and Sander Gribling. Quantum Speedups for Linear Programming via Interior Point Methods. *SIAM Journal on Computing*, 55(1):93–134, 2026. URL https://doi.org/10.1137/25M1736098.

Simon Apers, Sander Gribling, and Aaron Sidford. On computing approximate Lewis weights. *arXiv:2404.02881*, 2024.

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017. doi: 10.1287/moor.2016.0817.

J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Mathematica*, 162:73 – 141, 1989. doi: 10.1007/BF02392835. URL https://doi.org/10.1007/BF02392835.

Kenneth L. Clarkson and David P. Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):1–45, 2017.

Michael B. Cohen and Richard Peng. ℓ_p row sampling by Lewis weights. In *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*, pages 183–192, 2015.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1):3475–3506, 2012.

Maryam Fazel, Yin Tat Lee, Swati Padmanabhan, and Aaron Sidford. Computing Lewis weights to high precision. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2723–2742, 2022. doi: 10.1137/1.9781611977073.107. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611977073.107.

John Fritz. Extremum problems with inequalities as subsidiary conditions. In *Studies and Essays Presented to R. Courant on his 60th Birthday*. Interscience Publishers, New York, 1948.

Jean-Jacques Godeme, Jalal Fadili, Xavier Buet, Myriam Zerrad, Michel Lequime, and Claude Amra. Provable phase retrieval with mirror descent. *SIAM Journal on Imaging Sciences*, 16(3):1106–1141, 2023.

Arun Jambulapati, Yang P. Liu, and Aaron Sidford. Improved iteration complexities for overconstrained p-norm regression. In *Proceedings of the fifty-fourth annual ACM symposium on Theory of Computing*. ACM, 2022.

Leonid G. Khachiyan. Rounding of polytopes in the real number model of computation. *Mathematics of Operations Research*, 21(2):307–320, 1996.

Puya Latafat, Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Adaptive proximal algorithms for convex optimization under local lipschitz continuity of the gradient. *Mathematical Programming*, 213:433–471, 2025. doi: 10.1007/s10107-024-02143-7.

Yin Tat Lee. *Faster Algorithms for Convex and Combinatorial Optimization*. PhD thesis, Massachusetts Institute of Technology, 2016.

Yin Tat Lee and Aaron Sidford. Solving linear programs with $\sqrt{\text{rank}}$ linear system solves. *arXiv preprint arXiv:1910.08033*, 2019.

D. Lewis. Finite dimensional subspaces of L_p. *Studia Mathematica*, 63(2):207–212, 1978. URL http://eudml.org/doc/218208.

Yen-Huan Li, Carlos A. Riofrio, and Volkan Cevher. A general convergence result for mirror descent with armijo line search, 2018. arXiv:1805.12232.

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21–es, 2007.

Aaron Sidford and Kevin Tian. Coordinate methods for accelerating ℓ_∞ regression and faster approximate maximum flow. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 922–933, 2018. doi: 10.1109/FOCS.2018.00091.

Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *Proceedings of the fortieth annual ACM symposium on Theory of Computing*, pages 563–568, 2008.

Michel Talagrand. Embedding subspaces of L_1 into ℓ_1^N. *Proceedings of the American Mathematical Society*, 108(2):363–369, 1990.

Michael J. Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization, 2008.

Jan van den Brand, Yin Tat Lee, Yang P. Liu, Thatchaphol Saranurak, Aaron Sidford, Zhao Song, and Di Wang. Minimum cost flows, mdps, and ℓ_1-regression in nearly linear time for dense instances. In *Proceedings of the fifty-third annual ACM symposium on Theory of Computing*. ACM, 2021.

P. Wojtaszczyk. *Banach Spaces for Analysts*. Cambridge University Press, 1991.

David P. Woodruff and Taisuke Yasuda. Online lewis weight sampling. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4622–4666, 2023. doi: 10.1137/1.9781611977554.ch175. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611977554.ch175.

Appendix A. Technical lemmas

Lemma 20 *Let $h : \mathcal{S}^n_{\succ 0} \to \mathbb{R}$ be defined as $h(M) = -\operatorname{logdet}(M)$. Then, for any $M_1, M_2 \in \mathcal{S}^n_{\succ 0}$ satisfying $D_h(M_2, M_1) \le \varepsilon$ for some $\varepsilon \le 1/10$, we have*

$$(1 - 2\sqrt{\varepsilon})M_1 \preceq M_2 \preceq (1 + 2\sqrt{\varepsilon})M_1.$$

Proof Denote $\Delta \coloneqq M_1^{-1/2} M_2 M_1^{-1/2} \succ 0$. Then

$$D_h(M_2, M_1) = h(M_2) - h(M_1) - \langle \nabla h(M_1), M_2 - M_1 \rangle = \operatorname{Tr}(\Delta) - \log\det\Delta - n.$$

Hence, letting $\lambda_1, \dots, \lambda_n$ be the eigenvalues of Δ and $\phi(\lambda) \coloneqq \lambda - \log\lambda - 1$, we have

$$D_h(M_2, M_1) = \sum_{i \in [n]} \phi(\lambda_i).$$

Note that the function ϕ is convex on $(0, \infty)$, with a unique minimizer at $\lambda = 1$ and $\phi(1) = 0$. Hence, $\phi(\lambda) \ge 0$ for all $\lambda > 0$, which gives $\phi(\lambda_i) \le D_h(M_2, M_1) \le \varepsilon \le 1/10$ and thus $|\lambda_i - 1| \le 2\sqrt{\varepsilon}$ for all i. Therefore,

$$(1 - 2\sqrt{\varepsilon})I \preceq \Delta \preceq (1 + 2\sqrt{\varepsilon})I.$$

Conjugating by $M_1^{1/2}$ yields

$$(1 - 2\sqrt{\varepsilon})M_1 \preceq M_2 \preceq (1 + 2\sqrt{\varepsilon})M_1,$$

which completes the proof. ■

Lemma 21 *Suppose $x, y \in \mathbb{R}^m_{\ge 0}$ and $\delta > 0$ are such that $y \le (1+\delta)x$ entrywise and $\|x\|_1 \le (1+\delta)\|y\|_1$. Then $\|x - y\|_1 \le 3\delta\|y\|_1$.*

Proof The proof follows from writing $x - y = (x - \frac{1}{1+\delta}y) - \frac{\delta}{1+\delta}y$ and applying the triangle inequality:

$$\begin{aligned}\|x - y\|_1 &\le \sum_{i \in [m]} \left| x_i - \frac{1}{1+\delta} y_i \right| + \frac{\delta}{1+\delta} \sum_{i \in [m]} |y_i| = \sum_{i \in [m]} x_i - \frac{1}{1+\delta} y_i + \frac{\delta}{1+\delta} \sum_{i \in [m]} y_i \\ &= \|x\|_1 - \frac{1}{1+\delta}\|y\|_1 + \frac{\delta}{1+\delta}\|y\|_1.\end{aligned}$$

Finally, using $\|x\|_1 \le (1+\delta)\|y\|_1$ we obtain $\|x - y\|_1 \le \left(\frac{(1+\delta)^2 - 1 + \delta}{1+\delta}\right)\|y\|_1 \le 3\delta\|y\|_1$. ■

Lemma 22 $\operatorname{tr}[\hat{B} U \hat{B}^\top] \le \|U\|_1$ *for any full column rank $B \in \mathbb{R}^{m \times n}$ and PSD symmetric matrix $U \in \mathbb{R}^{n \times n}$, where $\hat{B} \coloneqq B(B^\top B)^{-1/2}$.*

Proof

$$\mathrm{tr}[\hat{B}U\hat{B}^\top] = \mathrm{tr}[\hat{B}U^{1/2}U^{1/2}\hat{B}^\top] = \mathrm{tr}[U^{1/2}\hat{B}^\top\hat{B}U^{1/2}]\,.$$

Note that

$$\hat{B}^\top\hat{B} = (B^\top B)^{-1/2}B^\top B(B^\top B)^{-1/2} \preceq I,$$

which leads to

$$\mathrm{tr}[\hat{B}U\hat{B}^\top] \leq \mathrm{tr}[U] \leq \|U\|_1.$$

■

Lemma 23 *For any $\zeta > 0$ and two full-rank PSD matrices M_1 and M_2 satisfying*

$$\|M_1^{-1/2}(M_2 - M_1)M_1^{-1/2}\|_1 \leq \zeta \leq \frac{1}{2}, \tag{21}$$

we have

$$\|M_2^{-1/2}(M_1 - M_2)M_2^{-1/2}\|_1 \leq \frac{\zeta}{1-\zeta}.$$

Proof Let $N := M_1^{-1/2}M_2M_1^{-1/2}$ so that (21) is equivalent to the statement that $\|N - I\|_1 \leq \zeta \leq 1/2$. Note that

$$\|M_2^{-1/2}(M_1 - M_2)M_2^{-1/2}\|_1 = \|N^{-1} - I\|_1 = \|(N-I)(N)^{-1}\|_1 \leq \|N - I\|_1\|N^{-1}\|_\infty\,.$$

However, since $\|N - I\|_\infty \leq \|N - I\|_1 \leq \zeta$ we know that every eigenvalue of N is between $1-\zeta$ and $1+\zeta$. Consequently, $\|N^{-1}\|_\infty \leq (1-\zeta)^{-1}$ yielding the result. ■

Lemma 24 *For any symmetric matrix M satisfying $\|M\|_2 \leq 1/2$, we have*

$$\left\| (I+M)^{-1} - I \right\|_1 \leq 2\|M\|_1.$$

Proof We use $\lambda_1, \ldots, \lambda_n$ to denote the eigenvalues of M. Then we have

$$\left\|(I+M)^{-1} - I\right\|_1 = \sum_{i\in[n]}\left|\frac{1}{1+\lambda_i} - 1\right| \leq 2\sum_{i\in[n]}|\lambda_i| = 2\|M\|_1,$$

where the inequality is due to the fact that for each λ_i we have $|\lambda_i| \leq \|M\|_1 \leq 1/2$. ■

Appendix B. Deferred proofs of Section 2

Lemma 25 (Three-Point Property, Tseng (2008)) *Let $\varphi\colon \mathcal{C} \to \mathbb{R}$ be convex. Given $z \in \mathbb{R}^d$, let $z^+ := \arg\min_{x\in\mathcal{C}}\{\varphi(x) + D_h(x,z)\}$. Then,*

$$\varphi(x) + D_h(x,z) \;\geq\; \varphi(z^+) + D_h(z^+,z) + D_h(x,z^+), \quad \forall x \in \mathcal{C}.$$

Proof of Lemma 11 By the L-locally relative smoothness condition, for any iteration t, we have

$$f(x^{(t+1)}) \leq f(x^{(t)}) + \langle \nabla f(x^{(t)}), x^{(t+1)} - x^{(t)}\rangle + LD_h(x^{(t+1)}, x^{(t)}).$$

Applying Lemma 25 with $\varphi(x) \coloneqq \langle \nabla f(x^{(t)}), x - x^{(t)}\rangle$ and using the fact that $x^{(t+1)} = \operatorname{argmin}_{x\in\mathcal{C}}$ $\left\{\varphi(x) + LD_h(x, x^{(t)})\right\}$, we obtain that for any $x \in \mathcal{C}$,

$$\begin{aligned}\langle \nabla f(x^{(t)}), x^{(t+1)} - x^{(t)}\rangle \leq \langle \nabla f(x^{(t)}), x - x^{(t)}\rangle + LD_h(x, x^{(t)}) \\ - LD_h(x^{(t+1)}, x^{(t)}) - LD_h(x, x^{(t+1)}).\end{aligned}$$

Therefore,

$$\begin{aligned}f(x^{(t+1)}) &\leq f(x^{(t)}) + \langle \nabla f(x^{(t)}), x - x^{(t)}\rangle + LD_h(x, x^{(t)}) - LD_h(x, x^{(t+1)}) \\ &\leq f(x) + (L-\mu)D_h(x, x^{(t)}) - LD_h(x, x^{(t+1)}),\end{aligned} \tag{22}$$

or equivalently,

$$D_h(x, x^{(t+1)}) \leq \left(1 - \frac{\mu}{L}\right) D_h(x, x^{(t)}) + \frac{1}{L}(f(x) - f(x^{(t+1)})).$$

■

Proof of Proposition 10. By Lemma 11, for each iteration t and any $x \in \mathcal{C}$ we have

$$D_h(x, x^{(t+1)}) \leq \left(1 - \frac{\mu}{L}\right) D_h(x, x^{(t)}) + \frac{1}{L}(f(x) - f(x^{(t+1)})).$$

For iteration t this yields

$$D_h(x, x^{(t)}) \leq \left(1 - \frac{\mu}{L}\right)^t D_h(x, x^{(0)}) + \frac{1}{L}\sum_{k\in[t]} \left(1 - \frac{\mu}{L}\right)^{t-k} (f(x) - f(x^{(k)})).$$

Substituting $x = x^*$ into the inequality above gives

$$D_h(x^*, x^{(t)}) \leq \left(1 - \frac{\mu}{L}\right)^t D_h(x^*, x^{(0)}),$$

and

$$\begin{aligned}\sum_{k\in[t]} \left(1 - \frac{\mu}{L}\right)^{t-k} (f(x^{(k)}) - f(x^*)) &\leq L\left(1 - \frac{\mu}{L}\right)^t D_h(x^*, x^{(0)}) - D_h(x^*, x^{(t)}) \\ &\leq L\left(1 - \frac{\mu}{L}\right)^t D_h(x^*, x^{(0)}),\end{aligned}$$

where we used that $f(x^*) - f(x^{(k)}) \leq 0$ for all k, and that $D_h(x^*, x^{(t)}) \geq 0$ since h is convex. Substituting $x = x^{(t)}$ in (22), we obtain $f(x^{(t+1)}) \leq f(x^{(t)}) - LD_h(x, x^{(t+1)}) \leq f(x^{(t)})$, which gives

$$\begin{aligned}\sum_{k\in[t]} \left(1 - \frac{\mu}{L}\right)^{t-k} (f(x^{(k)}) - f(x^*)) &\geq \sum_{k\in[t]} \left(1 - \frac{\mu}{L}\right)^{t-k} (f(x^{(t)}) - f(x^*)) \\ &= \frac{L}{\mu}\left(1 - \left(1 - \frac{\mu}{L}\right)^t\right)(f(x^{(t)}) - f(x^*)),\end{aligned}$$

and therefore

$$f(x^{(t)}) - f(x^*) \leq \frac{\mu(1-\mu/L)^t}{1-(1-\mu/L)^t} \cdot D_h(x^*, x^{(0)}).$$

■

Appendix C. Deferred proofs of Section 3

Proof of Lemma 12 Observe that

$$\min_{v>0} \mathcal{F}_{\text{vec}}(v) = \min_{v>0} \max_{M\succ 0} \left[\log\det(M) + n - \operatorname{Tr}(MA^\top VA) + \frac{1}{1+\alpha_p}\sum_{i\in[m]} v_i^{1+\alpha_p}\right] \tag{23}$$

Define

$$\Phi(v, M) \coloneqq \log\det(M) - \operatorname{Tr}(MA^\top VA) + \frac{1}{1+\alpha_p}\sum_{i\in[m]} v_i^{1+\alpha_p}.$$

Then, $\Phi(v, M)$ is convex with respect to $v \in \mathbb{R}^m_{>0}$ and concave with respect to $M \in \mathbb{R}^{n\times n}_{\succ 0}$. Moreover, $\mathcal{F}_{\text{vec}}$ diverges to $+\infty$ whenever any coordinate $v_i \to 0$ or $v_i \to \infty$. Therefore, $\mathcal{F}_{\text{vec}}$ admits a finite minimizer, and both the minimization over v and the maximization over M in (23) may be restricted to compact convex subsets without changing their values. Applying Sion's minimax theorem on these restricted domains then yields

$$\min_{v>0} \mathcal{F}_{\text{vec}}(v) = \max_{M\succ 0} \min_{v>0} \left[\log\det(M) + n - \operatorname{Tr}(MA^\top VA) + \frac{1}{1+\alpha_p}\sum_{i\in[m]} v_i^{1+\alpha_p}\right]$$

Furthermore, for any $M \in \mathbb{R}^{n\times n}_{\succ 0}$,

$$\begin{aligned}\min_{v>0}\Phi(v, M) &= \log\det(M) + \sum_{i\in[m]} \inf_{v_i>0}\left(\frac{1}{1+\alpha_p} v_i^{1+\alpha_p} - v_i\, a_i^\top M a_i\right) \\ &= \log\det(M) - \frac{2}{p}\sum_{i\in[m]} (a_i^\top M a_i)^{p/2} = -\mathcal{F}_{\text{mat}}(M),\end{aligned}$$

which gives

$$\min_{M\succ 0} \mathcal{F}_{\text{mat}}(M) = n - \min_{v>0} \mathcal{F}_{\text{vec}}(v).$$

■

Proof of Lemma 13 For any $M \in \mathbb{R}^{d\times d}_{\succ 0}$ and $H \in \mathbb{R}^{d\times d}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \log\det(M+tH) = \mathrm{Tr}(M^{-1}H) = \langle M^{-1}, H\rangle,$$

and since $p/2 - 1 = \frac{p-2}{2} = \beta_p$

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} (a_i^\top (M+tH)a_i)^{p/2} = \frac{p}{2}(a_i^\top M a_i)^{\beta_p} a_i^\top H a_i$$

for any $i \in [m]$. Hence,

$$\frac{\mathrm{d}\mathcal{F}_{\mathrm{mat}}(M+tH)}{\mathrm{d}t}\Big|_{t=0} = \Big\langle -M^{-1} + \sum_{i\in[m]} (a_i^\top M a_i)^{\beta_p} a_i a_i^\top, H\Big\rangle,$$

which gives

$$\nabla\mathcal{F}_{\mathrm{mat}}(M) = -M^{-1} + \sum_{i\in[m]} (a_i^\top M a_i)^{\beta_p} a_i a_i^\top.$$

Similarly, for any $M \in \mathbb{R}^{d\times d}_{\succ 0}$ and $H \in \mathbb{R}^{d\times d}$ we have

$$\lim_{t\to 0} \frac{\nabla\mathcal{F}_{\mathrm{mat}}(M+tK) - \nabla\mathcal{F}_{\mathrm{mat}}(M)}{t} = M^{-1}KM^{-1} + \frac{1}{\alpha_p}\sum_{i\in[m]} (a_i^\top M a_i)^{\beta_p - 1}(a_i^\top K a_i)\, a_i a_i^\top,$$

which gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \big\langle \nabla F(M+tK), H\big\rangle = \big\langle M^{-1}KM^{-1}, H\big\rangle + \frac{1}{\alpha_p}\sum_{i\in[m]} (a_i^\top M a_i)^{\beta_p-1}\langle a_i a_i^\top, H\rangle\langle a_i a_i^\top, K\rangle,$$

for all $M \in \mathbb{R}^{d\times d}_{\succ 0}$ and $H, K \in \mathbb{R}^{d\times d}$, which implies

$$\nabla^2\mathcal{F}_{\mathrm{mat}}(M) = M^{-1}\otimes M^{-1} + \frac{1}{\alpha_p}\sum_{i\in[m]} (a_i^\top M a_i)^{\beta_p-1} a_i a_i^\top \otimes a_i a_i^\top.$$

Note that $\nabla^2\mathcal{F}_{\mathrm{mat}}(M) \succ 0$. Hence, $\mathcal{F}_{\mathrm{mat}}$ is strictly convex and has a unique minimizer M_* satisfying $\nabla\mathcal{F}_{\mathrm{mat}}(M_*) = 0$, or equivalently,

$$M_*^{-1} = \sum_{i\in[m]} (a_i^\top M_* a_i)^{\beta_p} a_i a_i^\top.$$

Let $u \in \mathbb{R}^m_{>0}$ be the vector with coordinates $[u_*]_i = (a_i^\top M_* a_i)^{\beta_p}$. Then, we have $M_* = \big(AUA\big)^{-1}$ and

$$[u]_i = \big(a_i^\top \big(AUA\big)^{-1} a_i\big)^{\beta_p},$$

showing that $u = v_*$. ■

Algorithm 2: High-precision algorithm using $\mathcal{F}_{\text{vec}}$

Input: non-degenerate $A \in \mathbb{R}^{m\times n}$, $p > 2$, accuracy $\varepsilon \in (0, \frac{1}{4}]$

Set $\hat{\varepsilon} = \frac{\varepsilon}{\bar{\beta}_p n}\min\left\{\frac{1}{96(p-2)^2(4p-7)^2}, \frac{1}{50}\right\}$, $\Delta = \min\left\{\frac{\hat{\varepsilon}^3 n}{384\bar{\alpha}_p}, \frac{\hat{\varepsilon}^2\alpha_p^3}{27\times 10^3}\right\}$, $L = 32p\bar{\beta}_p$, $T = 2L\max\left\{\ln\left(\frac{4m}{\Delta}\right), 4\ln\left(\frac{5}{\alpha_p\hat{\varepsilon}}\right)\right\}$, and $v_i^{(0)} = 1$ for all $i \in [m]$.

for $t = 0, 1, \ldots, T-1$ **do**

  $v_i^{(t+1)} \leftarrow \left(1 + \frac{\rho_i(v^{(t)})-1}{L}\right)^{1/\alpha_p} v_i^{(t)}, \quad \forall i \in [m]$

end

Return $\widehat{w} \in \mathbb{R}^m_{>0}$, where $\widehat{w}_i = (a_i^\top (A^\top V^{(T)} A)^{-1} a_i)^{1+1/\alpha_p}$.

Appendix D. A relative smoothness algorithm using $\mathcal{F}_{\text{vec}}$

In this section, we present an algorithm that computes ε-estimates of Lewis weights by approximately solving (5) via the locally relatively smooth gradient descent framework in Section 2. Throughout this section, we set $\overline{\rho} = 4\bar{\beta}_p$, and define $r(v) \coloneqq \frac{1}{1+\alpha_p}\sum_{i\in[m]} v_i^{1+\alpha_p}$.

Theorem 7 *For $p > 2$, Algorithm 2 outputs an ε-estimate of $\sigma_p(A)$ in $O(p^2\log(mp^2\alpha_p/\varepsilon))$ iterations. Each iteration computes the leverage scores of DA of some diagonal matrix D.*

To prove Theorem 7, we first present several properties of $\mathcal{F}_{\text{vec}}$ in Section D.1. We then show in Section D.2 that the function value gap becomes sufficiently small after half of the iterations. Finally, in Section D.3, we establish that the final iterate $v^{(T)}$ of Algorithm 2 gives a one-sided $\hat{\varepsilon}$-approximate Lewis weight vector, which can be converted into an ε-estimate using Theorem 6. We present Algorithm 2 with the parameter ε as input since the main purpose of Algorithm 2 is to compute ε-estimates of $\sigma_p(A)$, but we point out that the one-sided $\hat{\varepsilon}$-approximation of Theorem 4 holds for any $\hat{\varepsilon} \in (0, \frac{1}{4}]$.

D.1. Properties of $\mathcal{F}_{\text{vec}}$

Here we present several properties of $\mathcal{F}_{\text{vec}}$, including explicit formulas for its gradient and Hessian, an upper bound on its function value gap for any $v \in \mathbb{R}^m_{\geq 0}$, and the fact that it is convex and locally smooth relative to r. We will use the projection matrix $P(v) \coloneqq V^{1/2}A(A^\top VA)^{-1}A^\top V^{1/2}$, and write $P(v)^{(2)}$ for the Schur product (entry-wise product) of $P(v)$ with itself.

Lemma 26 (Gradient and Hessian, Lemma 3 of Fazel et al. (2022)) *For any $v \in \mathbb{R}^m_{>0}$, the gradient and Hessian of $\mathcal{F}_{\text{vec}}$ have the following expressions:*

$$\left[\nabla\mathcal{F}_{\text{vec}}(v)\right]_i = v_i^{-1}\cdot(v_i^{1+\alpha_p} - \sigma_i(v)) \text{ and}$$
$$\nabla^2\mathcal{F}_{\text{vec}}(v) = V^{-1}P(v)^{(2)}V^{-1} + \alpha_p V^{\alpha_p - 1}.$$

Lemma 27 (Lemma 6 of Fazel et al. (2022)) *For any $v \in \mathbb{R}^m_{>0}$, we have*

$$\mathcal{F}_{\text{vec}}(v) - \mathcal{F}_{\text{vec}}(v_*) \geq \frac{1}{6\bar{\alpha}_p}\sum_{i\in[m]} v_i^{1+\alpha_p}\cdot\frac{(\rho_i(v)-1)^2}{\rho_i(v)+1}.$$

Lemma 28 *$\mathcal{F}_{\text{vec}}$ is 1-strongly convex relative to $r(v) \coloneqq \frac{1}{1+\alpha_p}\sum_{i\in[m]} v_i^{1+\alpha_p}$. Moreover,*

$$\nabla^2 \mathcal{F}_{\text{vec}}(v) \preceq \left(1 + \frac{\rho_{\max}(v)}{\alpha_p}\right) \nabla^2 r(v), \quad \forall v \in \mathbb{R}^m_{>0}.$$

Proof The gradient and Hessian of r satisfy

$$[\nabla r(v)]_i = v_i^{\alpha_p}, \qquad \nabla^2 r(v) = \alpha_p \cdot V^{\alpha_p - 1}.$$

Note that

$$\mathbf{0} \preceq \nabla^2 \left(-\operatorname{logdet}(A^\top V A)\right) = V^{-1} P(v)^{(2)} V^{-1} \preceq V^{-1}\mathbf{\Sigma} V^{-1} = \operatorname{Diag}(V^{\alpha_p - 1}\rho),$$

where we denote $\mathbf{\Sigma} \coloneqq \operatorname{Diag}(\sigma(v))$. Then by Lemma 26, we can conclude that

$$\nabla^2 r(v) \preceq \nabla^2 \mathcal{F}_{\text{vec}}(v) \preceq \left(1 + \frac{\rho_{\max}(v)}{\alpha_p}\right) \nabla^2 r(v)\,,$$

which shows that $\mathcal{F}_{\text{vec}}$ is 1-strongly convex relative to r. ■

D.2. Function value decrease in Algorithm 2

Here we show that the value of $\mathcal{F}_{\text{vec}}(v^{(t)}) - \mathcal{F}_{\text{vec}}(v_*)$ is at most Δ after $t \geq T/2$ iterations in Algorithm 2.

Lemma 29 *For any iteration t of Algorithm 2, we have*

$$v^{(t+1)} = \operatorname*{argmin}_{v\in\mathbb{R}^m_{>0}} \left\{\mathcal{F}_{\text{vec}}(v^{(t)}) + \left\langle \nabla \mathcal{F}_{\text{vec}}(v^{(t)}), v - v^{(t)}\right\rangle + L D_r(v, v^{(t)})\right\}. \tag{24}$$

Proof Since the function

$$\mathcal{F}_{\text{vec}}(v^{(t)}) + \langle \nabla \mathcal{F}_{\text{vec}}(v^{(t)}), v - v^{(t)}\rangle + L D_r(v, v^{(t)})$$

is convex, it has one unique minimizer. Given that

$$\begin{aligned}\nabla \left(\mathcal{F}_{\text{vec}}(v^{(t)}) + \left\langle \nabla \mathcal{F}_{\text{vec}}(v^{(t)}), v - v^{(t)}\right\rangle + L D_r(v, v^{(t)})\right)\Big|_{v=v^{(t+1)}} \\ = L\nabla r(v^{(t+1)}) + \nabla \mathcal{F}_{\text{vec}}(v^{(t)}) - L\nabla r(v^{(t)}) = 0,\end{aligned}$$

we can conclude that

$$v^{(t+1)} = \operatorname*{argmin}_{v\in\mathbb{R}^m_{>0}} \left\{\mathcal{F}_{\text{vec}}(v^{(t)}) + \left\langle \nabla \mathcal{F}_{\text{vec}}(v^{(t)}), v - v^{(t)}\right\rangle + L D_r(v, v^{(t)})\right\}.$$

■

The following two lemmas establish that $\mathcal{F}_{\text{vec}}$ is relatively smooth with respect to r between each pair of consecutive iterates $v^{(t)}$ and $v^{(t+1)}$ (and are analogs of Lemma 17 and 18 for $\mathcal{F}_{\text{vec}}$).

Lemma 30 *Let $L \geq 4\bar{\beta}_p$. For any iteration t in Algorithm 2, if $\rho_{\max}(v^{(t)}) \leq 4\bar{\beta}_p$, then $\rho_{\max}(v^{(t+1)}) \leq 4\bar{\beta}_p$. Consequently, $\rho_{\max}(v^{(t)}) \leq 4\bar{\beta}_p$ for all iterations t.*

Proof The proof strategy is similar to Lemma 17. First, note that $\rho_{\max}(v^{(0)}) \leq 1 \leq 4\bar{\beta}_p$. We then show that for any t satisfying $\rho_{\max}(v^{(t)}) \leq 4\bar{\beta}_p$, it also holds that $\rho_{\max}(v^{(t+1)}) \leq 4\bar{\beta}_p$. In particular, by the update formula in Line 3, $v_i^{(t+1)} \geq \left(1 - \frac{1}{L}\right)^{1/\alpha_p} v_i^{(t)} = \left(1 - \frac{1}{L}\right)^{\beta_p} v_i^{(t)}$, which leads to

$$\begin{aligned}
\rho_i(v^{(t+1)}) &= \left(\frac{v_i^{(t)}}{v_i^{(t+1)}}\right)^{\alpha_p} \cdot \frac{a_i^\top (A^\top V^{(t+1)} A)^{-1} a_i}{(v_i^{(t)})^{\alpha_p}} \\
&\leq \left(1 - \frac{1}{L}\right)^{-\beta_p} \left(\frac{v_i^{(t)}}{v_i^{(t+1)}}\right)^{\alpha_p} \cdot \frac{a_i^\top (A^\top V^{(t)} A)^{-1} a_i}{(v_i^{(t)})^{\alpha_p}} \\
&= \left(1 - \frac{1}{L}\right)^{-\beta_p} \left(1 + \frac{\rho_i(v^{(t)}) - 1}{L}\right)^{-1} \cdot \rho_i(v^{(t)}),
\end{aligned} \tag{25}$$

and the remainder of the proof proceeds analogously to proof of Lemma 17. ■

Lemma 31 *For any iteration t in Algorithm 2 and any $\lambda \in [0, 1]$, if $\rho_{\max}(v^{(t)}) \leq 4\bar{\beta}_p$ and $L \geq 4\bar{\beta}_p$, then*

$$\rho_{\max}((1 - \lambda)v^{(t)} + \lambda v^{(t+1)}) \leq 16\bar{\beta}_p.$$

Proof Let $u = (1-\lambda)v^{(t)} + \lambda v^{(t+1)}$. By the update formula in Line 3, we have $u_i = \left(1 + \lambda\left(\left(1 + \frac{\rho_i(v^{(t)})-1}{L}\right)^{1/\alpha_p} - 1\right)\right)$ for all $i \in [m]$, and therefore

$$u_i \geq \left(1 + \lambda\left(\left(1 - \frac{1}{L}\right)^{1/\alpha_p} - 1\right)\right) v_i^{(t)}.$$

Since $\beta_p = 1/\alpha_p$,

$$\rho_i(u) = \left(\frac{v_i^{(t)}}{u_i}\right)^{\alpha_p} \cdot \frac{a_i^\top (A^\top U A)^{-1} a_i}{(v_i^{(t)})^{\alpha_p}} \tag{26}$$

$$\leq \left(1 + \lambda\left(\left(1 - \frac{1}{L}\right)^{\beta_p} - 1\right)\right)^{-1} \left(\frac{v_i^{(t)}}{u_i}\right)^{\alpha_p} \cdot \frac{a_i^\top (A^\top V^{(t)} A)^{-1} a_i}{(v_i^{(t)})^{\alpha_p}} \tag{27}$$

$$= \left(1 + \lambda\left(\left(1 - \frac{1}{L}\right)^{\beta_p} - 1\right)\right)^{-1} \left(1 + \lambda\left(1 + \frac{\rho_i(v^{(t)}) - 1}{L}\right)^{\beta_p} - 1\right)^{-\alpha_p} \cdot \rho_i(v^{(t)}). \tag{28}$$

We bound the two multiplicative factors in (28) separately. For the first factor, we have

$$\left(1 + \lambda\left(\left(1 - \frac{1}{L}\right)^{\beta_p} - 1\right)\right)^{-1} \leq \left(1 + \left(\left(1 - \frac{1}{L}\right)^{\beta_p} - 1\right)\right)^{-1} \leq 2, \tag{29}$$

where the first inequality uses that $\left(1-\frac{1}{L}\right)^{\beta_p}-1\le 0$, and the last inequality uses that $\left(1-\frac{1}{L}\right)^{\beta_p}\ge \frac{1}{2}$ since $L\ge 4\bar{\beta}_p$. For the second factor in (28), we distinguish two cases: $\rho_i(v^{(t)})\le 1$ or $\rho_i(v^{(t)})>1$. On the one hand, when $\rho_i(v^{(t)})>1$, this second factor is at most 1. On the other hand, when $\rho_i(v^{(t)})\le 1$, we have $\left(1+\frac{\rho_i(v^{(t)})-1}{L}\right)^{\beta_p}-1\le 0$ and therefore this second factor increases when λ increases. This shows that

$$\left(1+\lambda\left(\left(1+\frac{\rho_i(v^{(t)})-1}{L}\right)^{\beta_p}-1\right)\right)^{-\alpha_p}\le\left(1+\left(\left(1-\frac{1}{L}\right)^{\beta_p}-1\right)\right)^{-\alpha_p},$$

which is at most $2^{\alpha_p}\le 2$ by (29) when $\alpha_p\le 1$, and is at most

$$\left(1-\frac{\beta_p}{L}\right)^{-\alpha_p}\le\left(1-\frac{1}{L}\right)^{-1}\le 2$$

when $\alpha_p>1$. Together, this shows that $\rho_i(u)\le 4\rho_i(v^{(t)})$ for each $i\in[m]$, which gives $\rho_{\max}(u)\le 4\rho_{\max}(v^{(t)})\le 16\bar{\beta}_p$. ■

Proposition 32 *For any iteration $t\ge T/2$ in Algorithm 2, we have $\mathcal{F}_{\text{vec}}(v^{(t)})-\mathcal{F}_{\text{vec}}(v_*)\le\Delta$.*

Proof Given our choice of L, Lemma 30 establishes that $\rho_{\max}(v^{(t)})\le 4\bar{\beta}_p$ for any iteration t of Algorithm 2. Then, by Lemma 28 and Lemma 31, it follows that $\mathcal{F}$ is $(1+16p\bar{\beta}_p)\le L$-smooth relative to r between any two consecutive iterates $v^{(t)}$ and $v^{(t+1)}$, and is 1-strongly convex relative to r. Hence, applying Proposition 10 gives

$$\mathcal{F}_{\text{vec}}(v^{(t)})-\mathcal{F}_{\text{vec}}(v_*)\le\frac{(1-1/L)^t}{1-(1-1/L)^t}D_r(v_*,v^{(0)})\le\Delta, \tag{30}$$

where we used the fact that

$$D_r(v_*,v^{(0)})=r(v_*)-r(v^{(0)})-\langle\nabla r(v^{(0)}),v_*-v^{(0)}\rangle\le 2m$$

since $[v_*]_i\le 1$ for any $i\in[m]$. ■

Remark 33 *By Lemma 30 and the initialization, $\rho_{\max}(v^{(t)})\le 4\bar{\beta}_p$ for every iteration t in Algorithm 2. Therefore, for every coordinate $i\in[m]$,*

$$\left|\frac{v_i^{(t+1)}}{v_i^{(t)}}-1\right|=\left|\left(1+\frac{\rho_i(v^{(t)})-1}{L}\right)^{1/\alpha_p}-1\right|\le\frac{1}{2}.$$

Consequently, the diagonal scaling $D^{(t)}=(V^{(t)})^{1/2}$ used in each leverage score computation satisfies

$$\sqrt{\frac{1}{2}}\,D_{ii}^{(t)}\le D_{ii}^{(t+1)}\le\sqrt{\frac{3}{2}}\,D_{ii}^{(t)},\quad\forall i\in[m].$$

D.3. One-sidedness property of the last iterate

Next, we show that the last iterate $v^{(T)}$ of Algorithm 2 provides one-sided approximate Lewis weights. To facilitate this analysis, for any $i, j \in [m]$, we define γ_{ij} to be the angle between the vectors $(A^\top VA)^{-1/2} v_i^{1/2} a_i$ and $(A^\top VA)^{-1/2} v_j^{1/2} a_j$, i.e.,

$$\gamma_{ij} \coloneqq \arccos\left(\frac{a_i^\top v_i^{1/2}(A^\top VA)^{-1} v_j^{1/2} a_j}{\|(A^\top VA)^{-1/2} v_i^{1/2} a_i\| \cdot \|(A^\top VA)^{-1/2} v_j^{1/2} a_j\|}\right).$$

The next three lemmas show that, for any iteration t of Algorithm 2 at which $\mathcal{F}_{\text{vec}}(v^{(t)})$ is sufficiently small, the value of $|\rho_i(v^{(t)}) - 1|$ will decrease multiplicatively until it is below certain threshold.

Lemma 34 (Lemma 47 of Lee and Sidford (2019)) *For any vector $v \in \mathbb{R}^m_{>0}$ and $i \in [m]$,*

$$\sum_{j \in [m]} \sigma_j(v) \cdot \cos^2(\gamma_{ij}) = 1, \qquad \forall i \in [m].$$

Lemma 35 *Let v, v^+ be vectors in $\mathbb{R}^m_{>0}$ such that $\frac{v}{2} \le v^+ \le \frac{3v}{2}$ and define*

$$\theta_i \coloneqq \sum_{j \in [m]} |v_j^+ - v_j| \cdot a_j^\top (A^\top VA)^{-1} a_j \cos(\gamma_{ij})^2 = \sum_{j \in [m]} \left|\frac{v_j^+}{v_j} - 1\right| \cdot \sigma_j(v) \cos(\gamma_{ij})^2.$$

Then for any i we have

$$(1 - 3\theta_i)\left(\frac{v_i^+}{v_i}\right)^{-\alpha_p} \le \frac{\rho_i(v^+)}{\rho_i(v)} \le (1 + 3\theta_i)\left(\frac{v_i^+}{v_i}\right)^{-\alpha_p}.$$

Proof Note that

$$\rho_i(v^{(+)}) = \left[\frac{v_i^{(+)}}{v_i}\right]^{-\alpha_p} \frac{1}{[v_i]^{\alpha_p}} a_i^\top (A^\top V^{(+)} A)^{-1} a_i = \left[\frac{v_i^{(+)}}{v_i}\right]^{-\alpha_p} \frac{a_i^\top (A^\top V^{(+)} A)^{-1} a_i}{a_i^\top (A^\top VA)^{-1} a_i} \cdot \rho_i(v).$$

Hence, it suffices to bound $\frac{a_i^\top (A^\top V^{(+)} A)^{-1} a_i}{a_i^\top (A^\top VA)^{-1} a_i}$. Denote $\Delta \coloneqq A^\top (V^{(+)} - V) A$. Then,

$$\begin{aligned}\frac{a_i^\top (A^\top V^{(+)} A)^{-1} a_i}{a_i^\top (A^\top VA)^{-1} a_i} &= \frac{a_i^\top (A^\top VA + \Delta)^{-1} a_i}{a_i^\top (A^\top VA)^{-1} a_i} \\ &= \frac{a_i^\top (A^\top VA)^{-1/2} (I + \bar{\Delta})^{-1} (A^\top VA)^{-1/2} a_i}{a_i^\top (A^\top VA)^{-1} a_i},\end{aligned}$$

where

$$\bar{\Delta} = (A^\top VA)^{-1/2} \Delta (A^\top VA)^{-1/2} = (A^\top VA)^{-1/2} A^\top (V^+ - V) A (A^\top VA)^{-1/2}.$$

From the assumption $v/2 \le v^+$ it follows that $\bar{\Delta} \succeq -I/2$. The latter in turn implies

$$1 - \bar{\Delta} \preceq (I + \bar{\Delta})^{-1} \preceq 1 - \bar{\Delta} + 2\bar{\Delta}^2.$$

Therefore, we have the following chain of inequalities

$$\begin{aligned}\frac{a_i^\top (A^\top V A)^{-1/2}(1-\bar{\Delta})(A^\top V A)^{-1/2} a_i}{a_i^\top (A^\top V A)^{-1} a_i} &\le \frac{a_i^\top (A^\top V^{(+)} A)^{-1} a_i}{a_i^\top (A^\top V A)^{-1} a_i} \\ &\le \frac{a_i^\top (A^\top V A)^{-1/2}(1-\bar{\Delta}+2\bar{\Delta}^2)(A^\top V A)^{-1/2} a_i}{a_i^\top (A^\top V A)^{-1} a_i}.\end{aligned} \tag{31}$$

We proceed by separately bounding the terms that depend linearly and quadratically on $\bar{\Delta}$. First, for the term that depends linearly on $\bar{\Delta}$, we have

$$\begin{aligned}|a_i^\top (A^\top V A)^{-1/2} \bar{\Delta} (A^\top V A)^{-1/2} a_i| &= |a_i^\top (A^\top V A)^{-1} A^\top (V^+ - V) A (A^\top V A)^{-1} a_i| \\ &\le \sum_{j\in[m]} |v_j^+ - v_j| \left(a_i^\top (A^\top V A)^{-1} a_j\right)^2 \\ &\le a_i^\top (A^\top V A)^{-1} a_i \sum_{j\in[m]} |v_j^+ - v_j| a_j^\top (A^\top V A)^{-1} a_j \cos(\gamma_{ij})^2\end{aligned} \tag{32}$$

Second, for the term that depends quadratically on $\bar{\Delta}$, we have

$$\begin{aligned}\bar{\Delta}^2 &= (A^\top V A)^{-1/2} A^\top (V^+ - V) A (A^\top V A)^{-1} A^\top (V^+ - V) A (A^\top V A)^{-1/2} \\ &= (A^\top V A)^{-1/2} A^\top (V^+ - V) V^{-1/2} V^{1/2} A (A^\top V A)^{-1} \\ &\qquad \cdot A^\top V^{1/2} V^{-1/2} (V^+ - V) A (A^\top V A)^{-1/2} \\ &\preceq (A^\top V A)^{-1/2} A^\top (V^+ - V) V^{-1/2} V^{-1/2} (V^+ - V) A (A^\top V A)^{-1/2} \\ &= (A^\top V A)^{-1/2} A^\top (V^+ - V)^2 V^{-1} A (A^\top V A)^{-1/2}.\end{aligned}$$

This allows us to proceed as before and obtain

$$\begin{aligned}&|a_i^\top (A^\top V A)^{-1/2} \bar{\Delta}^2 (A^\top V A)^{-1/2} a_i| \\ &\quad\le |a_i^\top (A^\top V A)^{-1} A^\top (V^+ - V)^2 V^{-1} A (A^\top V A)^{-1} a_i| \\ &\quad\le \sum_{j\in[m]} \frac{|v_j^+ - v_j|^2}{v_j} \left(a_i^\top (A^\top V A)^{-1} a_j\right)^2 \\ &\quad\le a_i^\top (A^\top V A)^{-1} a_i \sum_{j\in[m]} \frac{|v_j^+ - v_j|^2}{v_j} a_j^\top (A^\top V A)^{-1} a_j \cos(\gamma_{ij})^2 \\ &\quad\le a_i^\top (A^\top V A)^{-1} a_i \sum_{j\in[m]} |v_j^+ - v_j| a_j^\top (A^\top V A)^{-1} a_j \cos(\gamma_{ij})^2,\end{aligned} \tag{33}$$

where the last inequality uses the assumption $|v_j^+ - v_j| \le v_j$. Combining the above estimates (32) and (33) with (31) concludes the proof. ■

Lemma 36 *For any iteration t in Algorithm 2 and any $i \in [m]$, we have*

$$\rho_i(v^{(t+1)}) \le \max\left\{1 + 15L\theta(v^{(t)}), \left(1 - \frac{\rho_i(v^{(t)}) - 1}{4L}\right)\rho_i(v^{(t)})\right\} \tag{34}$$

where

$$\theta(v) \coloneqq \frac{2}{\alpha_p L}\sqrt{\frac{30}{\alpha_p}(\mathcal{F}_{\text{vec}}(v) - \mathcal{F}_{\text{vec}}(v_*))}, \quad \forall v \in \mathbb{R}^m_{\geq 0}.$$

Proof By the choice of L and Lemma 31, for any $i \in [m]$ we have

$$\left|\frac{v_i^{(t+1)}}{v_i^{(t)}} - 1\right| = \left|\left(1 + \frac{\rho_i(v^{(t)}) - 1}{L}\right)^{1/\alpha_p} - 1\right| \leq \frac{1}{2}.$$

Then, invoking Lemma 35 gives

$$(1 - 3\theta_i)\left(\frac{v_i^{(t+1)}}{v_i^{(t)}}\right)^{-\alpha_p} \leq \frac{\rho_i(v^{(t+1)})}{\rho_i(v^{(t)})} \leq (1 + 3\theta_i)\left(\frac{v_i^{(t+1)}}{v_i^{(t)}}\right)^{-\alpha_p},$$

where

$$\begin{aligned}
\theta_i &= \sum_{j\in[m]} \left|\frac{v_j^{(t+1)}}{v_j^{(t)}} - 1\right| \cdot \sigma_j(v^{(t)})\cos(\gamma_{ij})^2 \\
&= \sum_{j\in[m]} \left|\left(1 + \frac{\rho_j(v^{(t)}) - 1}{L}\right)^{1/\alpha_p} - 1\right| \cdot \sigma_j(v^{(t)})\cos(\gamma_{ij})^2 \\
&\leq \frac{2}{\alpha_p L}\sum_{j\in[m]} |\rho_j(v^{(t)}) - 1| \cdot \sigma_j(v^{(t)})\cos(\gamma_{ij})^2,
\end{aligned}$$

where the last inequality uses the fact that

$$\frac{\rho_j(v^{(t)}) - 1}{L} \leq \frac{\rho_{\max}(v^{(t)}) - 1}{L} \leq \frac{\alpha_p}{4}.$$

By Cauchy-Schwartz inequality, we have

$$\begin{aligned}
&\sum_{j\in[m]} |\rho_j(v^{(t)}) - 1| \cdot \sigma_j(v^{(t)})\cos(\gamma_{ij})^2 \\
&\qquad \leq \sqrt{\sum_{j\in[m]} \sigma_j(v^{(t)})(\rho_j(v^{(t)}) - 1)^2} \cdot \sqrt{\sum_{j\in[m]} \sigma_j(v^{(t)})\cos^2(\gamma_{ij})} \\
&\qquad \leq \sqrt{\sum_{j\in[m]} \sigma_j(v^{(t)})(\rho_j(v^{(t)}) - 1)^2} \\
&\qquad \leq \sqrt{\rho_{\max}(v^{(t)})}\sqrt{\sum_{j\in[m]} [v_j^{(t)}]^{1+\alpha_p}(\rho_j(v^{(t)}) - 1)^2} \\
&\qquad \leq \sqrt{\frac{30}{\alpha_p}(\mathcal{F}_{\text{vec}}(v^{(t)}) - \mathcal{F}_{\text{vec}}(v_*))},
\end{aligned}$$

where the second inequality uses Lemma 34 and the last inequality uses Lemma 27. Hence, we have $\theta_i \le \theta(v^{(t)})$ for all $i \in [m]$, which leads to

$$\begin{aligned}\rho_i(v^{(t+1)}) &\le \frac{1+3\theta(v^{(t)})}{1+\frac{\rho_i(v^{(t)})-1}{L}} \cdot \rho_i(v^{(t)}) \\ &\le (1+3\theta(v^{(t)}))\left(1-\frac{\rho_i(v^{(t)})-1}{2L}\right)\rho_i(v^{(t)}). \end{aligned} \tag{35}$$

Note that the function $\phi(x) \coloneqq \left(1-\frac{x-1}{2L}\right)x$ is monotonically increasing in $[0,1]$. Thus for any i with $\rho_i(v^{(t)}) \le 1$, the value of (35) is at most $1+3\theta(v^{(t)})$. Otherwise, we have

$$\rho_i(v^{(t+1)}) \le \max\left\{1+15L\theta(v^{(t)}), \left(1-\frac{\rho_i(v^{(t)})-1}{4L}\right)\rho_i(v^{(t)})\right\}.$$

■

Lemma 37 *For any $0 < \beta_p \le 1$ and $L \ge 1$, let $\{\zeta^{(t)}\}_{t\ge 0}$ be a sequence satisfying*

$$\zeta^{(t+1)} \le \max\left\{1+\beta_p, \left(1-\frac{\zeta^{(t)}-1}{4L}\right)\zeta^{(t)}\right\}, \quad \forall t \in \mathbb{N} \tag{36}$$

with $0 < \zeta^{(0)} \le L$ and $L \ge 1$. Then, there exists a finite index $\bar{t} = \lceil 4L\ln(\max\{\zeta^{(0)}, 1+\beta_p\}/\beta_p)\rceil$ such that

$$\zeta^{(t)} \le 1+\beta_p, \quad \forall t \ge \bar{t}.$$

Proof First note that for every $t \ge 0$, if $\zeta^{(t)} \le 1+\beta_p$, we have

$$\zeta^{(t+1)} \le \max\left\{1+\beta_p, \left(1-\frac{\zeta^{(t)}-1}{4L}\right)\zeta^{(t)}\right\} \le 1+\beta_p.$$

Hence, it suffices to prove that $\zeta^{(\bar{t})} \le 1+\beta_p$.

Assume the contrary, i,e, there exists a sequence $\{\zeta^{(t)}\}_{t\ge 0}$ satisfying (36) with $\zeta^{(\bar{t})} > 1+\beta_p$. Then, for any $t < \bar{t}$ we have $\zeta^{(t)} > 1+\beta_p$, and thus

$$\zeta^{(t+1)} - 1 \le \left(1-\frac{\zeta^{(t)}-1}{4L}\right)\zeta^{(t)} - 1 \le \left(1-\frac{1}{4L}\right)(\zeta^{(t)}-1),$$

which leads to

$$\zeta^{(\bar{t})} \le 1+\left(1-\frac{1}{4L}\right)^{\bar{t}}(\zeta^{(0)}-1) \le 1+\exp\left(-\frac{\bar{t}}{4L}\right)\zeta^{(0)} \le 1+\beta_p,$$

contradiction. Therefore, we can conclude that we have $\zeta^{(t)} \le 1+\beta_p$ for any sequence $\{\zeta^{(t)}\}_{t\ge 0}$ satisfying (36) and any $t \ge \bar{t}$. ■

Lemma 38 *For any $0 < \varepsilon \le 1/2$ and any $v \in \mathbb{R}^m_{>0}$ satisfying $\rho_{\max}(v) \le 1+\varepsilon$ and $\mathcal{F}_{\text{vec}}(v) - \mathcal{F}_{\text{vec}}(v_*) \le \varepsilon^3 n/(384\bar{\alpha}_p)$, the vector $w = v^{1+\alpha_p}$ is a one-sided ε-approximation of $\sigma_p(A)$.*

Proof We directly have $\sigma(w^{\frac{1}{2}-\frac{1}{p}}) \le (1+\varepsilon)w$ since $\rho_{\max}(w^{\frac{1}{2}-\frac{1}{p}}) = \rho_{\max}(v) \le 1+\varepsilon$. As for the upper bound on $\|w\|_1$, we denote

$$\Gamma \coloneqq \{i \in [m] \mid \rho_i(v) \ge 1 - \varepsilon/4\}. \tag{37}$$

Then, $\|w\|_1 = \sum_{i\in\Gamma} w_i + \sum_{i\in[m]\setminus\Gamma} w_i$, where we have

$$\sum_{i\in\Gamma} w_i \le \frac{1}{1-\varepsilon/4} \sum_{i\in\Gamma} \sigma_i(w^{\frac{1}{2}-\frac{1}{p}}) \le \left(1+\frac{\varepsilon}{2}\right) \sum_{i\in[m]} \sigma_i(w^{\frac{1}{2}-\frac{1}{p}}) \le \left(1+\frac{\varepsilon}{2}\right) n \tag{38}$$

and

$$\begin{aligned}\sum_{i\in[m]\setminus\Gamma} w_i = \sum_{i\in[m]\setminus\Gamma} v_i^{1+\alpha_p} &\le \frac{32}{\varepsilon^2} \sum_{i\in[m]\setminus\Gamma} v_i^{1+\alpha_p} \frac{(\rho_i(v)-1)^2}{\rho_i(w)+1} \\ &\le \frac{192\bar{\alpha}_p}{\varepsilon^2}(\mathcal{F}_{\text{vec}}(v) - \mathcal{F}_{\text{vec}}(v_*)) \le \frac{\varepsilon n}{2},\end{aligned}$$

where the second inequality uses Lemma 27. Hence, we can conclude that $\|w\|_1 \le (1+\varepsilon)n$ and thus w are one-sided ε-approximate Lewis weights. ■

Theorem 4 *For $p > 2$, Algorithm 2 with parameter $\hat{\varepsilon}$ produces, after $T = O(p^2 \log(mp^2\alpha_p/\hat{\varepsilon}))$ iterations, a vector $w \coloneqq [v^{(T)}]^{1+\alpha_p}$ that is a one-sided $\hat{\varepsilon}$-approximation of $\sigma_p(A)$.*

Proof By Proposition 32, for any iteration $t \ge T/2$ we have $\mathcal{F}_{\text{vec}}(v^{(t)}) - \mathcal{F}_{\text{vec}}(v_*) \le \Delta$. Then, invoking Lemma 36 gives

$$\rho_i(v^{(t+1)}) \le \max\left\{1+\hat{\varepsilon}, \left(1 - \frac{\rho_i(v^{(t)})-1}{4L}\right)\rho_i(v^{(t)})\right\}$$

for any $i \in [m]$ and $t \ge T/2$. Therefore, by Lemma 37, we have $\rho_i(v^{(T)}) \le 1+\hat{\varepsilon}$ for any $i \in [m]$, or equivalently, $\rho_{\max}(v^{(T)}) \le 1+\hat{\varepsilon}$. Using Lemma 38, we can conclude that w is a one-sided $\hat{\varepsilon}$-approximation of $\sigma_p(A)$. ■

Theorem 7 can then be established by combining Theorem 4 and Theorem 6.

Appendix E. Conversion between different approximation guarantees

In this section, we first show how to convert one-sided approximations of $\sigma_p(A)$ to two-sided approximations in Section E.1, and then how to convert them to estimates of $\sigma_p(A)$ in Section E.2. In Section E.3, we show how to convert approximate minimizers of $\mathcal{F}_{\text{vec}}$ to two-sided approximations. Finally, in Section E.4, we present an improved analysis of a variant of the algorithm in Lee (2016); the full guarantee is stated in Theorem 49. Throughout this section, we denote $H \coloneqq A^\top W^{\frac{1}{2}-\frac{1}{p}} A$, $\widehat{H} \coloneqq A^\top \widehat{W}^{\frac{1}{2}-\frac{1}{p}} A$, and $\hat{\rho} \coloneqq \rho(\widehat{w}^{\frac{1}{2}-\frac{1}{p}}) \in \mathbb{R}^m_{>0}$.

E.1. From one-sided approximations to two-sided approximations

Here we establish the following result.

Theorem 5 *For $p \geq 2$, if w is a one-sided ε_{one}-approximation of $\sigma_p(A)$ and $\widehat{w} := \sigma(w^{\frac{1}{2}-\frac{1}{p}})^{\frac{p}{2}}/w^{\beta_p}$, then $\widehat{w}$ is a two-sided ε_{two}-approximation of $\sigma_p(A)$ for $\varepsilon_{\text{two}} = 3\bar{\beta}_p n\varepsilon_{\text{one}}(1+\varepsilon_{\text{one}})^{\bar{\beta}_p}$.*

We first prove a key lemma that allows us to compare quadratic forms associated with $\widehat{w}$ and w.

Lemma 39 *For any $w \in \mathbb{R}^m_{>0}$ and $\varepsilon_{\text{two}} := \left\|\widehat{w} - \sigma(w^{\frac{1}{2}-\frac{1}{p}})\right\|_1$, we have*

$$\|H^{-1/2}(\widehat{H}-H)H^{-1/2}\|_1 \leq \left\|\widehat{w} - \sigma(w^{\frac{1}{2}-\frac{1}{p}})\right\|_1 = \varepsilon_{\text{two}},$$

and consequently,

$$(1-\varepsilon_{\text{two}})H \preceq \widehat{H} \preceq (1+\varepsilon_{\text{two}})H.$$

Proof Let $\Delta := \widehat{w}^{1-\frac{2}{p}} - w^{1-\frac{2}{p}}$ and let $\Delta_+ := \max\{\Delta, \mathbf{0}\}$ and $\Delta_- := \max\{-\Delta, \mathbf{0}\}$ entrywise so that $\Delta_+, \Delta_- \in \mathbb{R}^m_{\geq 0}$, $\Delta = \Delta_+ - \Delta_-$, and $\|\Delta\|_1 = \|\Delta_+\|_1 + \|\Delta_-\|_1$. Then, we have

$$\left\|H^{-1/2}(\widehat{H}-H)H^{-1/2}\right\|_1 \leq \left\|H^{-1/2}\left(A^\top\Delta_+A\right)H^{-1/2}\right\|_1 + \left\|H^{-1/2}\left(A^\top\Delta_-A\right)H^{-1/2}\right\|_1 .$$

We then use the fact Δ_+ is positive semidefinite, and therefore so is $H^{-1/2}\left(A^\top\Delta_+A\right)H^{-1/2}$, to upper bound the spectral norm by the trace:

$$\begin{aligned}\left\|H^{-1/2}\left(A^\top\Delta_+A\right)H^{-1/2}\right\|_1 &= \operatorname{tr}\left[H^{-1/2}\left(A^\top\Delta_+A\right)H^{-1/2}\right] \\ &= \sum_{i\in[m]}[\Delta_+]_i\left[A(A^\top W^{\frac{1}{2}-\frac{1}{p}}A)^{-1}A^\top\right]_{ii} = \sum_{i\in[m]}[\Delta_+]_i\cdot\frac{\sigma_i(w^{\frac{1}{2}-\frac{1}{p}})}{w_i^{1-\frac{2}{p}}}.\end{aligned}$$

By symmetry, the same bound holds for Δ_-. Combining the two bounds yields that

$$\left\|H^{-1/2}(A^\top\Delta A)H^{-1/2}\right\|_1 \leq \sum_{i\in[m]}|\Delta_i|\cdot\frac{\sigma_i(w^{\frac{1}{2}-\frac{1}{p}})}{w_i^{1-\frac{2}{p}}}$$

Recalling that $\Delta = \widehat{w}^{1-\frac{2}{p}} - w^{1-\frac{2}{p}}$ we obtain the desired upper bound:

$$\sum_{i\in[m]}|\Delta_i|\cdot\frac{\sigma_i(w^{\frac{1}{2}-\frac{1}{p}})}{w_i^{1-\frac{2}{p}}} = \left\|\widehat{w} - \sigma(w^{\frac{1}{2}-\frac{1}{p}})\right\|_1 = \varepsilon_{\text{two}}\,.$$

Finally, we conclude that

$$\left\|H^{-1/2}(A^\top\Delta A)H^{-1/2}\right\|_2 \leq \left\|H^{-1/2}(A^\top\Delta A)H^{-1/2}\right\|_1 \leq \varepsilon_{\text{two}} \tag{39}$$

and thus $(1-\varepsilon_{\text{two}})H \preceq \widehat{H} \preceq (1+\varepsilon_{\text{two}})H$. ■

Next, we show how to bound the approximation factor $\varepsilon_{\text{two}} = \left\|\widehat{w} - \sigma(w^{\frac{1}{2}-\frac{1}{p}})\right\|_1$ when w is a one-sided approximation of $\sigma_p(A)$.

Lemma 40 *If $w \in \mathbb{R}^m_{>0}$ is a one-sided ε_{one}-approximate ℓ_p-Lewis weight of A for $p > 2$, then*

$$\left\|\widehat{w} - \sigma(w^{\frac{1}{2}-\frac{1}{p}})\right\|_1 \leq 3\bar{\beta}_p n \varepsilon_{\text{one}}(1+\varepsilon_{\text{one}})^{\bar{\beta}_p}.$$

Proof If $p \geq 4$ and $\beta_p \geq 1$, since $\widehat{w}$ and $\sigma(w^{\frac{1}{2}-\frac{1}{p}})$ are non-negative, it follows that

$$\begin{aligned}\left\|\widehat{w} - \sigma(w^{\frac{1}{2}-\frac{1}{p}})\right\|_1 &= \sum_{i\in[m]} \sigma_i(w^{\frac{1}{2}-\frac{1}{p}}) \left|\left(\frac{\sigma_i(w^{\frac{1}{2}-\frac{1}{p}})}{w_i}\right)^{\beta_p} - 1\right| \\ &\leq \beta_p(1+\varepsilon_{\text{one}})^{\beta_p}\|\sigma(w^{\frac{1}{2}-\frac{1}{p}}) - w\|_1\,,\end{aligned}$$

where in the last step we used the fact that

$$|1-x^c| = \left|\int_1^x c\cdot y^{c-1}\mathrm{d}y\right| \leq \left|\int_1^x c\cdot \max\{1,x\}^{c-1}\mathrm{d}y\right| = c\cdot\max\{1,x\}^{c-1}\cdot|1-x|$$

for any $x \geq 0$ and $c > 1$, and that $\sigma(w^{\frac{1}{2}-\frac{1}{p}}) \leq (1+\varepsilon_{\text{one}})w$. Otherwise, if $2 \leq p < 4$ and $0 < \beta_p < 1$, we have

$$\begin{aligned}\left\|\widehat{w} - \sigma(w^{\frac{1}{2}-\frac{1}{p}})\right\|_1 &= \sum_{i\in[m]} \sigma_i(w^{\frac{1}{2}-\frac{1}{p}}) \left|\left(\frac{\sigma_i(w^{\frac{1}{2}-\frac{1}{p}})}{w_i}\right)^{\beta_p} - 1\right| \\ &\leq \sum_{i\in[m]} \frac{\sigma_i(w^{\frac{1}{2}-\frac{1}{p}})}{w_i}\left|\sigma_i(w^{\frac{1}{2}-\frac{1}{p}}) - w_i\right| \leq (1+\varepsilon_{\text{one}})\|\sigma(w^{\frac{1}{2}-\frac{1}{p}}) - w\|_1.\end{aligned}$$

The result then follows from the fact that $\|w - \sigma(w^{\frac{1}{2}-\frac{1}{p}})\|_1 \leq 3\varepsilon_{\text{one}}\|\sigma(w)\|_1 = 3\varepsilon_{\text{one}}n$, by Lemma 21. ■

Now we are ready to prove Theorem 5.

Proof of Theorem 5 For any $i \in [m]$, we have

$$\frac{\sigma(\widehat{w}^{\frac{1}{2}-\frac{1}{p}})_i}{\widehat{w}_i} = \widehat{w}_i^{-\frac{2}{p}} \cdot \left[A(A^\top \widehat{W}^{1-\frac{2}{p}}A)^{-1}A^\top\right]_{ii} = \frac{\left[A(A^\top\widehat{W}^{1-\frac{2}{p}}A)^{-1}A^\top\right]_{ii}}{\left[A(A^\top W^{1-\frac{2}{p}}A)^{-1}A^\top\right]_{ii}}.$$

We wish to lower and upper bound this fraction by $1/(1+\varepsilon_{\text{two}})$ and $1/(1-\varepsilon_{\text{two}})$, respectively. For this it suffices to prove that

$$\frac{1}{1+\varepsilon_{\text{two}}}\left(A^\top W^{1-\frac{2}{p}}A\right)^{-1} \preceq \left(A^\top\widehat{W}^{1-\frac{2}{p}}A\right)^{-1} \preceq \frac{1}{1-\varepsilon_{\text{two}}}\left(A^\top W^{1-\frac{2}{p}}A\right)^{-1},$$

which is equivalent to showing that

$$(1-\varepsilon_{\text{two}})A^\top W^{1-\frac{2}{p}}A \preceq A^\top\widehat{W}^{1-\frac{2}{p}}A \preceq (1+\varepsilon_{\text{two}})A^\top W^{1-\frac{2}{p}}A.$$

By Lemma 39 and Lemma 40 this holds for $\varepsilon_{\text{two}} = \|\widehat{w} - \sigma(w^{\frac{1}{2}-\frac{1}{p}})\|_1 \leq 3\bar{\beta}_p n\varepsilon_{\text{one}}(1+\varepsilon_{\text{one}})^{\bar{\beta}_p}$. ■

E.2. From one-sided approximations to estimates of Lewis weights

Theorem 6 *For $p > 2$, suppose w is a one-sided ε_{one}-approximation of $\sigma_p(A)$ satisfying*

$$\varepsilon_{\text{one}} \le \frac{1}{\bar{\beta}_p n} \min\left\{ \frac{1}{96(p-2)^2(4p-7)^2}, \frac{1}{50} \right\}.$$

Define $\widehat{w} \in \mathbb{R}^m_{>0}$ by $\widehat{w}_i := \sigma_i(w^{\frac{1}{2}-\frac{1}{p}})^{\frac{p}{2}} / w_i^{\beta_p}$ for each $i \in [m]$. Then $\widehat{w}$ is an ε_{est}-estimate of $\sigma_p(A)$, where $\varepsilon_{\text{est}} = 2(p-2)(4p-7)\sqrt{6\bar{\beta}_p n \varepsilon_{\text{one}}}$.

Lemma 41 *Let $\widehat{\Sigma} = \operatorname{diag}(\sigma_i(\widehat{w}^{\frac{1}{2}-\frac{1}{p}}))$ and $\varepsilon_{\text{two}} = \|\widehat{w} - \sigma(w^{\frac{1}{2}-\frac{1}{p}})\|_1$. If $\varepsilon_{\text{two}} \le 1/8$ then $\|\ln \rho(\widehat{w}^{\frac{1}{2}-\frac{1}{p}})\|_{\widehat{\Sigma}} \le 4\varepsilon_{\text{two}}^{1/2}$.*

Proof

$$\rho_i(\widehat{w}^{\frac{1}{2}-\frac{1}{p}}) = \widehat{w}_i^{-\frac{2}{p}} a_i^\top (A^\top \widehat{W}^{1-\frac{2}{p}} A)^{-1} a_i = \frac{a_i^\top (A^\top \widehat{W}^{1-\frac{2}{p}} A)^{-1} a_i}{a_i^\top (A^\top W^{1-\frac{2}{p}} A)^{-1} a_i} = \frac{a_i^\top \widehat{H}^{-1} a_i}{a_i^\top H^{-1} a_i},$$

which leads to

$$\rho_i(\widehat{w}^{\frac{1}{2}-\frac{1}{p}})^{-1} = \frac{a_i^\top \widehat{H}^{-\frac{1}{2}} \widehat{H}^{\frac{1}{2}} H^{-1} \widehat{H}^{\frac{1}{2}} \widehat{H}^{-\frac{1}{2}} a_i}{a_i^\top \widehat{H}^{-1} a_i}$$

where

$$\widehat{H}^{\frac{1}{2}} H^{-1} \widehat{H}^{\frac{1}{2}} = \left(\widehat{H}^{-\frac{1}{2}} H \widehat{H}^{-\frac{1}{2}}\right)^{-1} = \left(I + \widehat{H}^{-\frac{1}{2}}(H - \widehat{H})\widehat{H}^{-\frac{1}{2}}\right)^{-1}.$$

By Lemma 39 and Lemma 23, we have

$$\left\| \widehat{H}^{-\frac{1}{2}}(H - \widehat{H})\widehat{H}^{-\frac{1}{2}} \right\|_1 \le 2\varepsilon_{\text{two}}.$$

Denote

$$\Delta := \left(I + \widehat{H}^{-\frac{1}{2}}(H - \widehat{H})\widehat{H}^{-\frac{1}{2}}\right)^{-1} - I.$$

Then, $\|\Delta\|_1 \le 2\varepsilon_{\text{two}}$ by Lemma 24. Define $\Delta_+ := \max\{\Delta, 0\}$ and $\Delta_- := \max\{-\Delta, 0\}$ entry-wise, and define a new positive semidefinite matrix $\bar{\Delta} := \Delta_+ - \Delta_-$ that satisfies

$$\|\bar{\Delta}\|_1 \le \|\Delta_+\|_1 + \|\Delta_-\|_1 \le 2\|\Delta\|_1 \le 4\varepsilon_{\text{two}}.$$

Moreover, we have

$$\left|\rho_i(\widehat{w}^{\frac{1}{2}-\frac{1}{p}})^{-1} - 1\right| \le \frac{\|\widehat{H}^{-\frac{1}{2}} a_i\|_{\bar{\Delta}}^2}{\|\widehat{H}^{-\frac{1}{2}} a_i\|_2^2} \le \|\bar{\Delta}\|_2 \le 4\varepsilon_{\text{two}} \le \frac{1}{2},$$

which leads to

$$|\ln(\rho_i(\widehat{w}^{\frac{1}{2}-\frac{1}{p}}))| \le \frac{2\|\widehat{H}^{-\frac{1}{2}} a_i\|_{\bar{\Delta}}^2}{\|\widehat{H}^{-\frac{1}{2}} a_i\|_2^2} = \frac{2\|\widehat{H}^{-\frac{1}{2}} \widehat{w}_i^{\frac{1}{2}-\frac{1}{p}} a_i\|_{\bar{\Delta}}^2}{\|\widehat{H}^{-\frac{1}{2}} \widehat{w}_i^{\frac{1}{2}-\frac{1}{p}} a_i\|_2^2},$$

and

$$\|\ln \rho(\widehat{w}^{\frac{1}{2}-\frac{1}{p}})\|_{\widehat{\Sigma}}^2 = \sum_{i\in[m]} \sigma_i(\widehat{w}^{\frac{1}{2}-\frac{1}{p}}) |\ln(\rho_i(\widehat{w}^{\frac{1}{2}-\frac{1}{p}}))|^2 \le 4 \sum_{i\in[m]} \sigma_i(\widehat{w}^{\frac{1}{2}-\frac{1}{p}}) \cdot \frac{\|\widehat{H}^{-\frac{1}{2}} \widehat{w}_i^{\frac{1}{2}-\frac{1}{p}} a_i\|_{\bar{\Delta}}^4}{\|\widehat{H}^{-\frac{1}{2}} \widehat{w}_i^{\frac{1}{2}-\frac{1}{p}} a_i\|_2^4}.$$

Denote $\hat{A} = \widehat{W}^{\frac{1}{2}-\frac{1}{p}} A \widehat{H}^{-\frac{1}{2}}$. Then by Lemma 22,

$$\begin{aligned}\|\ln \rho(\widehat{w}^{\frac{1}{2}-\frac{1}{p}})\|_{\Sigma}^2 &\le 4 \sum_{i\in[m]} \sigma_i(\widehat{w}^{\frac{1}{2}-\frac{1}{p}}) \cdot \left(\frac{\hat{a}_i^\top \bar{\Delta} \hat{a}_i}{\hat{a}_i^\top \hat{a}_i}\right)^2 \le 4 \sum_{i\in[m]} \sigma_i(\widehat{w}^{\frac{1}{2}-\frac{1}{p}}) \cdot \left(\frac{\hat{a}_i^\top \bar{\Delta} \hat{a}_i}{\hat{a}_i^\top \hat{a}_i}\right) \\ &= 4\mathrm{tr}[\hat{A}\bar{\Delta}\hat{A}^\top] \le \|\bar{\Delta}\|_1 \le 16\varepsilon_{\mathrm{two}}\end{aligned}$$

here we use that $\left(\frac{\hat{a}_i^\top \bar{\Delta} \hat{a}_i}{\hat{a}_i^\top \hat{a}_i}\right) \le 1$ and that $\sigma_i(\widehat{w}^{\frac{1}{2}-\frac{1}{p}}) = \hat{a}_i^\top \hat{a}_i$. ■

Lemma 42 (Lemma 14 and Claim 1 of Fazel et al. (2022)) *Consider $\hat{v} := \widehat{w}^{1-\frac{2}{p}}$ which satisfies $|\rho_i(\hat{v}) - 1| \le 1/2$ for any i, define*

$$\hat{v}(t) = \operatorname*{argmin}_{v\in\mathbb{R}^m_{>0}} f_t(v) := -\log\det\left(A^\top V A\right) + \frac{1}{1+\alpha_p}\sum_{i=1}^m \rho_i^t(\hat{v}) v_i^{1+\alpha_p}, \quad \forall t \in [0,1]. \tag{40}$$

Then, we have $\hat{v}(1) = \hat{v}$, $\hat{v}(0) = v_*$, and

$$\left\| \frac{\mathrm{d}}{\mathrm{d}t} \ln\left(\frac{\hat{v}(t)}{\hat{v}(1)}\right)\right\|_\infty \le \frac{\|\ln \rho(\hat{v})\|_\infty}{\alpha_p} + \frac{\|\ln \rho(\hat{v})\|_{\Sigma(\hat{v}(t))}}{\alpha_p^2},$$

Lemma 43 *For any $\xi \in \mathbb{R}^m_{>0}$ that satisfies $\|\ln \xi\|_\infty = \gamma \le 1/4$, we have*

$$\left\| \frac{\sigma(\mathrm{diag}(\xi)A)}{\sigma(A)}\right\|_\infty \le 1 + 8\gamma.$$

Proof Given that $\|\ln \xi\|_\infty = \gamma \le 1/4$, we have $\frac{1}{1+2\gamma} \le \xi_i \le 1 + 2\gamma$ for all $i \in [m]$. Then for each i, we have

$$\begin{aligned}\sigma_i(\xi A) &= \xi_i^2 a_i^\top (A^\top \mathrm{diag}(\xi)^2 A)^{-1} a_i \\ &\le (1+2\gamma)^2 a_i^\top \left(\frac{A^\top A}{(1+2\gamma)^2}\right)^{-1} a_i \le (1+2\gamma)^4 \sigma_i(A) \le (1+8\gamma)\sigma_i(A).\end{aligned}$$

■

Proof of Theorem 6 By Theorem 5, we have that $\widehat{w}$ is an $\varepsilon_{\mathrm{two}}$-two-sided approximation of $\sigma_p(A)$ for some $\varepsilon_{\mathrm{two}}$ satisfying

$$\varepsilon_{\mathrm{two}} \le 3\bar{\beta}_p n \varepsilon_{\mathrm{one}}(1+\varepsilon_{\mathrm{one}})^{\bar{\beta}_p} \le 6\bar{\beta}_p n \varepsilon_{\mathrm{one}} \le \frac{1}{8} \tag{41}$$

by our choice of ε_{one}. Assume $\hat{v} = \widehat{w}^{1-\frac{2}{p}}$ satisfies $\|\ln(\hat{v}/v_*)\|_\infty \le \frac{1}{4}$, which will be justified later. Consider the vector function $\hat{v}(t)$ defined in (40), by Theorem 5 and Lemma 42, we have

$$\left\|\frac{\mathrm{d}}{\mathrm{d}t}\ln\left(\frac{\hat{v}(t)}{\hat{v}(1)}\right)\right\|_\infty \le \frac{\|\ln\hat{\rho}\|_\infty}{\alpha_p} + \frac{\|\ln\hat{\rho}\|_{\Sigma(\hat{v}(t))}}{\alpha_p^2} \le \frac{2\varepsilon_{\text{two}}}{\alpha_p} + \frac{\|\ln\hat{\rho}\|_{\Sigma(\hat{v}(t))}}{\alpha_p^2}.$$

By Lemma 43,

$$\|\ln\hat{\rho}\|_{\Sigma(\hat{v}(t))} \le \|\ln\hat{\rho}\|_{\Sigma(\hat{v})} \cdot \left\|\frac{\sigma(\hat{v}(t))}{\sigma(\hat{v}(1))}\right\|_\infty \le \|\ln\hat{\rho}\|_{\Sigma(\hat{v})}\left(1 + 8\left\|\ln\left(\frac{\hat{v}(t)}{\hat{v}(1)}\right)\right\|_\infty\right).$$

Moreover, since

$$\left\|\ln\left(\frac{\hat{v}(t)}{\hat{v}(1)}\right)\right\|_\infty \le \left\|\ln\left(\frac{\hat{v}(0)}{\hat{v}(1)}\right)\right\|_\infty \le \frac{1}{4},$$

and $\|\ln\hat{\rho}\|_{\Sigma(\hat{v})} \le 4\varepsilon_{\text{two}}$ by Lemma 41, we obtain

$$\begin{aligned}\left\|\frac{\mathrm{d}}{\mathrm{d}t}\ln\left(\frac{\hat{v}(t)}{\hat{v}(1)}\right)\right\|_\infty &\le \frac{p-2}{2}\varepsilon_{\text{two}} + 4\varepsilon_{\text{two}}^{1/2}\left(\frac{p-2}{2}\right)^2\left(1 + 8\left\|\ln\left(\frac{\hat{v}(t)}{\hat{v}(1)}\right)\right\|_\infty\right)\\ &\le (p-2)\varepsilon_{\text{two}} + 4(p-2)^2\varepsilon_{\text{two}}^{1/2} \le \frac{1}{4}.\end{aligned}$$

Integrating over $t \in [0,1]$ gives

$$\left\|\ln\left(\frac{\hat{v}(t)}{\hat{v}(1)}\right)\right\|_\infty \le (p-2)\,(4p-7)\,\varepsilon_{\text{two}}^{1/2} \le (p-2)\,(4p-7)\sqrt{6\bar{\beta}_p n\varepsilon_{\text{one}}} \le \frac{1}{4}, \quad \forall t \in [0,1],$$

where the second inequality uses (41). This verifies the assumption made at the beginning of the proof. Since $\hat{v}(0) = v_*$, it follows that

$$\exp\big(-(p-2)\,(4p-7)\,\varepsilon_{\text{two}}^{1/2}\big)v_* \le \hat{v} \le \exp\big((p-2)\,(4p-7)\,\varepsilon_{\text{two}}^{1/2}\big)v_*$$

and thus $(1-\varepsilon_{\text{est}})\sigma_p(A) \le \widehat{w} \le (1+\varepsilon_{\text{est}})\sigma_p(A)$. ■

E.3. From approximate minimizers of $\mathcal{F}_{\text{vec}}$ to two-sided approximations

Here we establish the following result.

Theorem 8 *For $p > 2$ and $\varepsilon \le \min\{\frac{1}{1000}, \frac{1}{50\bar{\alpha}_p}\}$, suppose $v \in \mathbb{R}^m_{>0}$ satisfies $\mathcal{F}_{\text{vec}}(v) - \mathcal{F}_{\text{vec}}(v_*) \le \varepsilon^3$ and $\rho_{\max}(v) \le 1+\varepsilon$. Define $\widetilde{w} \in \mathbb{R}^m_{>0}$ coordinatewise by setting $\widetilde{w}_i = (\sigma_i(v)/v_i)^{1+1/\alpha_p}$ if $\rho_i(v) \le 1-\varepsilon$, and $\widetilde{w}_i = v_i^{1+\alpha_p}$ otherwise. Then $\widetilde{w}$ is a two-sided $50\max\{\alpha_p, 1\}\varepsilon$-approximation of $\sigma_p(A)$.*

For comparison, prior work established the following conversion from approximate optimality to estimates. Ours achieves two-sided approximations instead of estimates, but without polynomial overhead in terms of the dimension.

Lemma 44 (Lemma 1 of Fazel et al. (2022)) *For any $v \in \mathbb{R}^m_{>0}$ satisfying $\rho_{\max}(v) \leq 1 + \alpha_p$ and $\mathcal{F}_{\text{vec}}(v) - \mathcal{F}_{\text{vec}}(v_*) \leq \tilde{\varepsilon}$ with*

$$\tilde{\varepsilon} = \frac{\alpha_p^8 \varepsilon^4}{\left(25m(\sqrt{n} + \alpha_p)(\alpha_p + \alpha_p^{-1})\right)^4},$$

Then, the vector $\widehat{w}$ defined as $\widehat{w}_i = \left(a_i^\top (A^\top V A)^{-1} a_i\right)^{1+1/\alpha_p}$, is an ε-estimate of $\sigma_p(A)$.

We first prove a key lemma that allows us to compare quadratic forms associated with $\widetilde{w}$ and w.

Lemma 45 *Let $v \in \mathbb{R}^m_{>0}$, $S \subseteq [m]$, and define*

$$\widetilde{v}_i = \begin{cases} (\sigma_i(v)/v_i)^{1/\alpha_p} & \text{if } i \in S \\ v_i & \text{otherwise.} \end{cases}$$

Then

$$(1-\delta) A^\top V A \preceq A^\top \widetilde{V} A \preceq (1+\delta) A^\top V A,$$

where $\delta \coloneqq \sum_{i \in S} |\widetilde{v}_i^{1+\alpha_p} - \sigma_i(v)|$.

Proof Let $\Delta := \widetilde{v} - v$ and let $\Delta_+ := \max\{\Delta, \mathbf{0}\}$ and $\Delta_- := \max\{-\Delta, \mathbf{0}\}$ entrywise so that $\Delta_+, \Delta_- \in \mathbb{R}^m_{\geq 0}$, $\Delta = \Delta_+ - \Delta_-$. Let $H = A^\top V A$ and $\widetilde{H} = A^\top \widetilde{V} A$. Then, we have

$$\left\| H^{-1/2}(\widetilde{H} - H) H^{-1/2} \right\|_1 \leq \left\| H^{-1/2}\left(A^\top \Delta_+ A\right) H^{-1/2} \right\|_1 + \left\| H^{-1/2}\left(A^\top \Delta_- A\right) H^{-1/2} \right\|_1 .$$

We then use the fact Δ_+ is positive semidefinite, and therefore so is $H^{-1/2}\left(A^\top \Delta_+ A\right) H^{-1/2}$, to upper bound the spectral norm by the trace:

$$\left\| H^{-1/2}\left(A^\top \Delta_+ A\right) H^{-1/2} \right\|_1 = \operatorname{tr}\left[H^{-1/2}\left(A^\top \Delta_+ A\right) H^{-1/2} \right] = \sum_{i \in [m]} [\Delta_+]_i \left[A (A^\top V A)^{-1} A^\top \right]_{ii}$$

$$= \sum_{i \in [m]} [\Delta_+]_i \cdot \frac{\sigma_i(v)}{v_i} .$$

By symmetry the same bound holds for Δ_-. Combining the two bounds yields that

$$\left\| H^{-1/2}(A^\top \Delta A) H^{-1/2} \right\|_1 \leq \sum_{i \in [m]} |\Delta_i| \cdot \frac{\sigma_i(v)}{v_i}$$

Recalling that $\Delta = \widetilde{v} - v$ we obtain the desired upper bound:

$$\sum_{i \in [m]} |\Delta_i| \cdot \frac{\sigma_i(v)}{v_i} = \sum_{i \in S} |\widetilde{v}_i^{1+\alpha_p} - \sigma_i(v)| = \delta.$$

Finally, we conclude that

$$\left\| H^{-1/2}(A^\top \Delta A) H^{-1/2} \right\|_2 \leq \left\| H^{-1/2}(A^\top \Delta A) H^{-1/2} \right\|_1 \leq \delta \tag{42}$$

and thus $(1-\delta) H \preceq \widetilde{H} \preceq (1+\delta) H$. ■

Lemma 46 *For any $0 < \varepsilon < 1$ and any $v \in \mathbb{R}^m_{>0}$ we have*

$$\sum_{i\in[m]:\rho_i(v)\le 1-\varepsilon} v_i^{1+\alpha_p} \le \frac{6\bar{\alpha}_p\left(\mathcal{F}_{\mathrm{vec}}(v) - \mathcal{F}_{\mathrm{vec}}(v_*)\right)\left(1+\rho_{\max}(v)\right)}{\varepsilon^2}.$$

Proof This follows immediately from Lemma 27. Indeed,

$$\sum_{i\in[m]:\rho_i(v)\le 1-\varepsilon} v_i^{1+\alpha_p} \le \sum_{i\in[m]:\rho_i(v)\le 1-\varepsilon} v_i^{1+\alpha_p}\frac{(\rho_i(v)-1)^2}{\varepsilon^2}\frac{\rho_{\max}(v)+1}{\rho_i(v)+1} \tag{43}$$

$$\le \frac{\rho_{\max}(v)+1}{\varepsilon^2}\sum_{i\in[m]} v_i^{1+\alpha_p}\frac{(\rho_i(v)-1)^2}{\rho_i(v)+1} \tag{44}$$

$$\le \frac{6\bar{\alpha}_p\left(\mathcal{F}_{\mathrm{vec}}(v) - \mathcal{F}_{\mathrm{vec}}(v_*)\right)\left(1+\rho_{\max}(v)\right)}{\varepsilon^2}, \tag{45}$$

where the last inequality uses Lemma 27. ■

Lemma 47 *Let $v \in \mathbb{R}^m_{>0}$ be such that $\mathcal{F}_{\mathrm{vec}}(v) - \mathcal{F}_{\mathrm{vec}}(v_*) \le \varepsilon^3$ and $\rho_{\max}(v) \le 1+\varepsilon$. Let $\widetilde{v}$ be defined as follows*

$$\widetilde{v}_i = \begin{cases} (\sigma_i(v)/v_i)^{1/\alpha_p} & \text{if } \rho_i(v) \le 1-\varepsilon \\ v_i & \text{otherwise.} \end{cases}$$

Then $\sum_{i\in[m]:\rho_i(v)\le 1-\varepsilon} |\widetilde{v}_i^{1+\alpha_p} - \sigma_i(v)| \le 12\bar{\alpha}_p(1+\varepsilon)\varepsilon$

Proof Apply the triangle inequality. Observe that for each i for which $\rho_i(v) \le 1-\varepsilon$ we have: $\widetilde{v}_i^{1+\alpha_p} \le v_i^{1+\alpha_p}$ since $(\sigma_i(v)/v_i)^{1/\alpha_p} = v_i\rho_i(v)^{1/\alpha_p}$, and $\sigma_i(v) \le v_i^{1+\alpha_p}$. Finally, apply Lemma 46. ■

Proof of Theorem 8 We first note that by Lemma 45 and Lemma 47 we have

$$(1-\delta)A^\top V A \preceq A^\top \widetilde{V} A \preceq (1+\delta)A^\top V A$$

for $\delta = 24\bar{\alpha}_p\varepsilon$. This shows that

$$\frac{1}{1+\delta}\left(A^\top V A\right)^+ \preceq \left(A^\top \widetilde{V} A\right)^+ \preceq \frac{1}{1-\delta}\left(A^\top V A\right)^+, \tag{46}$$

We now bound $\rho_i(\widetilde{v})$. We distinguish two cases: $\rho_i(v) < 1-\varepsilon$ and $\rho_i(v) \ge 1-\varepsilon$. First, when $\rho_i(v) < 1-\varepsilon$ we have $\widetilde{v}_i = (\sigma_i(v)/v_i)^{1/\alpha_p}$, and therefore

$$\rho_i(\widetilde{v}) = \widetilde{v}_i^{-\alpha_p}\cdot\left[A(A^\top\widetilde{V}A)^{-1}A^\top\right]_{ii} = \frac{\left[A(A^\top\widetilde{V}A)^+A^\top\right]_{ii}}{[A(A^\top V A)^+A^\top]_{ii}}.$$

(46) then shows that $\rho_i(\widetilde{v}) \in [(1+\delta)^{-1}, (1-\delta)^{-1}]$.

Second, when $\rho_i(v) \geq 1-\varepsilon$ we proceed as follows. We have $\widetilde{v}_i = v_i$ and therefore

$$\rho_i(\widetilde{v}) = v_i^{-\alpha_p} \cdot \left[A(A^\top \widetilde{V} A)^{-1} A^\top\right]_{ii} = \rho_i(v) \cdot \frac{\left[A(A^\top \widetilde{V} A)^{-1} A^\top\right]_{ii}}{[A(A^\top V A)^{-1} A^\top]_{ii}}.$$

(46) then shows that $\rho_i(\widetilde{v}) \in [(1+\delta)^{-1}\rho_i(v), (1-\delta)^{-1}\rho_i(v)] \subseteq [\frac{1-\varepsilon}{1+\delta}, \frac{1+\varepsilon}{1-\delta}]$.

Combining the two cases shows that $\widetilde{w} = \widetilde{v}^{1+\alpha_p}$ is a two-sided $\tilde{\varepsilon}$-approximation of $\sigma_p(A)$ for $\tilde{\varepsilon} = \frac{\delta+\varepsilon}{1-\varepsilon} = 25\bar{\alpha}_p\varepsilon/(1-\varepsilon) \leq 50\bar{\alpha}_p\varepsilon$. ■

E.4. Improved analysis of Lee's algorithm

Through Theorem 5, we can show that a variation of Lee's algorithm can be used to compute ε-estimates of $\sigma_p(A)$ using approximate leverage score computations at the expense of a $\mathrm{poly}(n,p)$-overhead in precision and a dimension dependent number of iterations, see Theorem 49.

We first prove a simple lemma that shows that two-sided approximation is "stable" with respect to a multiplicative change (i.e., if w is a two-sided approximation then so is its multiplicative approximation).

Lemma 48 *Let $\gamma \geq 1$. Let $w, \widetilde{w} \in \mathbb{R}^m_{>0}$ be such that $\gamma^{-1}\widetilde{w}_i \leq w_i \leq \gamma\widetilde{w}_i$ for all $i \in [m]$. Then*

$$\gamma^{-1}\frac{\sigma_i(\widetilde{W}^{\frac{1}{2}-\frac{1}{p}}A)}{\widetilde{w}_i} \leq \frac{\sigma_i(W^{\frac{1}{2}-\frac{1}{p}}A)}{w_i} \leq \gamma\frac{\sigma_i(\widetilde{W}^{\frac{1}{2}-\frac{1}{p}}A)}{\widetilde{w}_i}$$

Proof We first prove the first inequality. We have that

$$\begin{aligned}\frac{\sigma_i(\widetilde{W}^{\frac{1}{2}-\frac{1}{p}}A)}{\widetilde{w}_i} &= \widetilde{w}_i^{-\frac{2}{p}} \cdot \left[A(A^\top \widetilde{W}^{1-\frac{2}{p}}A)^+ A^\top\right]_{ii} \\ &\leq \gamma\, w_i^{-\frac{2}{p}} \cdot \left[A(A^\top W^{1-\frac{2}{p}}A)^+ A^\top\right]_{ii} = \gamma\frac{\sigma_i(W^{\frac{1}{2}-\frac{1}{p}}A)}{w_i}\end{aligned}$$

where the inequality uses $\widetilde{w}_i^{-\frac{2}{p}} \leq \gamma^{\frac{2}{p}} w_i^{-\frac{2}{p}}$ and $(A^\top \widetilde{W}^{1-\frac{2}{p}}A)^+ \preceq \gamma^{1-\frac{2}{p}}(A^\top W^{1-\frac{2}{p}}A)^+$. The second inequality of the lemma follows by exchanging the roles of w and $\widetilde{w}$. ■

We can now state our variation of Lee's algorithm and prove its correctness.

Theorem 49 (Approximate Lewis weights from approximate leverage scores) *Algorithm 3 outputs a two-sided ε-approximation of the ℓ_p-Lewis weights of A. Each iteration computes the leverage scores of DA of some diagonal matrix D to multiplicative accuracy $O(\varepsilon/(pn))$.*

Proof Steps 1.-4. of the algorithm correspond to Algorithm 6 by Lee (2016). In Theorem 5.3.4 of Lee (2016) it is shown that the resulting w satisfies $w_i/\sigma_i(W^{\frac{1}{2}-\frac{1}{p}}A) \geq \exp(-\varepsilon_1)$ and hence $\sigma_i(W^{\frac{1}{2}-\frac{1}{p}}A) \leq \exp(\varepsilon_1)w_i \leq (1+2\varepsilon_1)w_i$. Moreover, w is an average over $\varepsilon_1/4$-approximate leverage scores so that $\|w\|_1 \leq (1+\varepsilon_1/4)n$, and hence w is a one-sided $2\varepsilon_1$-approximation of

Algorithm 3: Two-sided Lewis weight approximation

Input: $A \in \mathbb{R}^{m\times n}$, $p \geq 2$, accuracy $\varepsilon > 0$

Let $w_i^{(1)} = n/m$ for all $i \in [m]$, $\varepsilon_1 = \varepsilon/(100pn)$, $\varepsilon_2 = \varepsilon/(3p)$, $T = \lceil 2\log(m/n)/\varepsilon_1 \rceil$;

for $k = 1, \ldots, T-1$ **do**

  Let $w^{(k+1)}$ be $\varepsilon_1/4$-estimates of $\sigma((W^{(k)})^{\frac{1}{2}-\frac{1}{p}}A)$;

end

Let $w = \frac{1}{T}\sum_{k\in[T]} w^{(k)}$ and s be ε_2-estimates of $\sigma(W^{\frac{1}{2}-\frac{1}{p}}A)$;

return $\widetilde{w}$ *with* $\widetilde{w}_i = w_i(s_i/w_i)^{\frac{p}{2}}$ *for all* $i \in [m]$

$\sigma_p(A)$. By Theorem 5 this implies the vector $\widehat{w} := \sigma(w^{\frac{1}{2}-\frac{1}{p}})^{\frac{p}{2}}/w^{\beta_p}$ is a two-sided Lewis weight approximation with approximation factor

$$6\bar{\beta}_p n\varepsilon_1(1+2\varepsilon_1)^{\bar{\beta}_p} n \leq \varepsilon/3$$

by our choice of ε_1. Finally, we use ε_2-estimates s of $\sigma(W^{\frac{1}{2}-\frac{1}{p}}A)$ to define $\widetilde{w}_i = w_i(s_i/w_i)^{\frac{p}{2}}$, so that

$$(1-\varepsilon_2)^{p/2}\widehat{w}_i \leq \widetilde{w}_i \leq (1+\varepsilon_2)^{p/2}\widehat{w}_i \leq \frac{1}{(1-\varepsilon_2)^{p/2}}\widehat{w}_i.$$

We can now apply Lemma 48 with $\gamma = 1/(1-\varepsilon_2)^{p/2} \leq 1+\varepsilon/3$ by our choice of ε_2. This implies that the $\widetilde{w}_i$'s are two-sided Lewis weight approximations satisfying

$$(1-\varepsilon/3)^2 \leq \frac{\sigma_i(\widetilde{W}^{\frac{1}{2}-\frac{1}{p}}A)}{\widetilde{w}_i} \leq (1+\varepsilon/3)^2.$$

Using that $(1-\varepsilon/3)^2 \geq 1-\varepsilon$ and $(1+\varepsilon/3)^2 \leq 1+\varepsilon$, this proves the claim. ■

Theorem 49 implies that we can obtain a two-sided ε-approximation of $\sigma_p(A)$ by iteratively computing $O(pn\log m/\varepsilon)$ many $O(\varepsilon/(pn))$-approximate leverage scores.