A Computing AFAC points

In this appendix we prove Theorem 4, which states that AFAC points can be computed in polynomial time.

We focus on the constrained optimization problem (P_{con}) throughout this appendix. We will later need the following lemma, which gives a simple bound on the magnitude of the Lagrange multipliers of (P_{con}).

Lemma 4. *Let (y, λ) be such that $\|\nabla_y L(y, \lambda)\| \leq \varepsilon_1$. If ϱ-LICQ holds at y, then $\|\lambda\| \leq \varrho^{-1}(\varepsilon_1 + \|\nabla f(y)\|)$.*

Proof. Let $J := \nabla h(y)$. Since $\nabla L(y, \lambda) = \nabla f(y) + J^T \lambda$, and J is full rank, then $\lambda = (J^\dagger)^T(\nabla L(y, \lambda) - \nabla f(y))$, where $J^\dagger$ is the pseudo-inverse of J. Hence $\|\lambda\| \leq \varrho^{-1}(\varepsilon_1 + \|\nabla f(y)\|)$. $\square$

A.1 The algorithm

Cartis et al. [14] proposed a method for computing q-th order critical points for $q \in \{1, 2, 3\}$. However, they use a nonstandard notion of criticality which is not easy to translate into our setting. We present here a slight modification of this algorithm that accommodates more general criticality conditions.

Consider the least squares functions

$$\nu(y) := \|h(y)\|^2, \qquad \mu(t, y) := (f(y) - t)^2 + \|h(y)\|^2.$$

We denote $\mu_t = \mu(t, \cdot)$ the function obtained by fixing the value of t. Algorithm 1 below is a variant of the method from [14]. It consists of two phases. The first phase attempts to find an approximately feasible solution through the unconstrained problem $\min_y \nu(y)$. If successful, the second phase minimizes f while preserving feasibility. To do so, it solves a sequence of problems $\min_y \mu(t_k, y)$, where the values $\{t_k\}_{k \geq 0}$ are decreasing.

Algorithm 1 Constrained optimization algorithm based on [14]

Input: Initial point $y_0 \in \mathbb{R}^n$, tolerances $\epsilon_0 \in \mathbb{R}_+$, $\boldsymbol{\epsilon} \in \mathbb{R}^q_+$, constant $\delta \in (0, 1)$.
Output: A point $y \in \mathbb{R}^n$ and a number $t \leq f(y)$.

PHASE I

$y_1 := \operatorname{local\,min}_y \nu(y)$ starting with y_0
$t_0 := f(y_1)$
if $\nu(y_1) > (\delta\epsilon_0)^2$ **then return** (y_1, t_0)

PHASE II

$t_1 := f(y_1) - (\epsilon_0^2 - \nu(y_1))^{1/2}$
for $k = 2, 3, 4, \ldots$ **do**
  $y_k := \operatorname{local\,min}_y \mu(t_{k-1}, y)$ starting with y_{k-1}
  if $\mu(t_{k-1}, y_k) < (\delta\epsilon_0)^2$ **then** ▷ case (a)
    $t_k := f(y_k) - (\epsilon_0^2 - \nu(y_k))^{1/2}$
    if $\boldsymbol{\chi}(\mu_{t_k}, y_k) \leq \boldsymbol{\epsilon}$ **then return** (y_k, t_k)
  if $\mu(t_{k-1}, y_k) \geq (\delta\epsilon_0)^2$ & $f(y_k) < t_{k-1}$ **then** ▷ case (b)
    $t_k := 2f(y_k) - t_{k-1}$
    if $\boldsymbol{\chi}(\mu_{t_k}, y_k) \leq \boldsymbol{\epsilon}$ **then return** (y_k, t_k)
  if $\mu(t_{k-1}, y_k) \geq (\delta\epsilon_0)^2$ & $f(y_k) \geq t_{k-1}$ **then** ▷ case (c)
    return (y_k, t_k), with $t_k := t_{k-1}$

Algorithm 1 relies on an *inner method* for solving the unconstrained problem $\min_y \psi(y)$, where ψ is either ν or $\mu_t = \mu(t, \cdot)$. Given $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_q) \in \mathbb{R}^q_+$, the inner method looks for a point y such that $\boldsymbol{\chi}(\psi, y) \leq \boldsymbol{\epsilon}$, for some *criticality measure* $\boldsymbol{\chi} = (\chi_1, \ldots, \chi_q)$. We assume that the j-th component $\chi_j(\psi, y)$ only involves derivatives $\{\nabla^d \psi(y)\}_{d \leq j}$ up to order j. For instance, the AC-criticality condition from (1) corresponds to the case

$$\boldsymbol{\chi}^{\text{AC}}(\psi, y) := \big(\|\nabla \psi(y)\|, -\min \operatorname{eig}(\nabla^2 \psi(y)) \big). \tag{10}$$

Given an initial point y^0 and tolerances $\boldsymbol{\epsilon} \in \mathbb{R}^q_+$, the inner method produces iterates $\{y^i\}_{i=1}^N$. We assume that the final point y^N achieves these tolerances and that the objective function decreases proportionately to N:

$$\boldsymbol{\chi}(\psi, y^N) \leq \boldsymbol{\epsilon} \qquad \text{and} \qquad \psi(y^0) - \psi(y^N) \geq N\, \kappa_\psi\, p(\boldsymbol{\epsilon}), \tag{11}$$

for some $\kappa_\psi > 0$ and some function p. Hence, the number of iterations N is proportional to $p(\boldsymbol{\epsilon})^{-1}$.

The next theorem provides guarantees for Algorithm 1. Our proof closely follows that of [14, Thm.4.5] but has the advantage that it applies to a general class of criticality measures, as opposed to [14], which relies on a particular nonstandard measure of criticality. However our complexity is larger than in [14] by a factor of ϵ_0^{-1}.

Theorem 8. *Assume that:*

- *The inner method satisfies* (11) *for the function ν with constant κ_ν.*
- *The inner method satisfies* (11) *for the function μ_t, and the constant κ_μ is independent of t.*
- *There exists $\beta > \epsilon_0$ and $f_{\text{low}} \in \mathbb{R}$ such that $f(y) \geq f_{\text{low}}$ for all $y \in \mathfrak{M}_\beta$, where $\mathfrak{M}_\beta := \{y : \|h(y)\| \leq \beta\}$.*

Then the total number of inner iterations made in Algorithm 1 is at most

$$p(\boldsymbol{\epsilon})^{-1} \left(\kappa_\nu^{-1}\, \nu(y_0) \;+\; \epsilon_0\, \kappa_\mu^{-1}\, (1-\delta)^{-1}\, (f(y_1) - f_{\text{low}} + \beta)\right), \tag{12}$$

and the algorithm returns a pair (y, t) such that:

$$\begin{aligned} &\textit{either} \qquad && t < f(y), \qquad && \|h(y)\| \leq \epsilon_0, \qquad && \boldsymbol{\chi}(\mu_t, y) \leq \boldsymbol{\epsilon}, \qquad && \text{(13a)} \\ &\textit{or} && t = f(y), && \|h(y)\| > \delta\epsilon_0, && \chi_1(\nu, y) \leq \epsilon_1. && \text{(13b)} \end{aligned}$$

A.2 Proof of Theorem 8

Let K be the number of outer iterations of Algorithm 1. Consider the sets of indices:

$$\begin{aligned} A &:= \{1\} \cup \{k : 2 \leq k \leq K \text{ and case (a) is applied}\}, \\ B &:= \{k : 2 \leq k \leq K \text{ and case (b) is applied}\}. \end{aligned}$$

The following lemma gives a few properties of Algorithm 1. Its proof is identical to [14, Lem.3.1].

Lemma 5. *If the algorithm reaches Phase II, then:*

$$\begin{aligned} &\nu(y_k) \leq \mu(t_k, y_k) \leq \epsilon_0^2, \qquad && 0 \leq f(y_k) - t_k \leq \epsilon_0, \qquad && \textit{for } k \geq 1, \qquad && (14) \\ &\mu(t_k, y_k) = \epsilon_0^2, && t_{k-1} - t_k \geq (1-\delta)\epsilon_0, && \textit{for } k \in A, && (15) \\ &\mu(t_k, y_k) = \mu(t_{k-1}, y_k), && t_{k-1} > t_k, && \textit{for } k \in B, && (16) \\ &\mu(t_k, y_k) \geq (\delta\epsilon_0)^2, && \boldsymbol{\chi}(\mu_{t_k}, y_k) \leq \boldsymbol{\epsilon}, && \textit{for } k = K. && (17) \end{aligned}$$

Let (y, t) be the output of Algorithm 1, and let us show (13). Assume first that the algorithm terminates in Phase I. Then y is a local minimum of ν, $\nu(y) > (\delta\epsilon_0)^2$, and $t = f(y)$. Hence (13b) holds. Assume now that the algorithm terminates in Phase II. By (14) and (17), we have

$$t \leq f(y), \qquad (\delta\epsilon_0)^2 \leq \mu_t(y) \leq \epsilon_0^2, \qquad \boldsymbol{\chi}(\mu_t, y) \leq \boldsymbol{\epsilon}.$$

If $f(y) < t$ then $\|h(y)\| \leq \sqrt{\mu_t(y)} \leq \epsilon_0$, so (13a) holds. Consider now the case that $f(y) = t$. Note that $\mu_t(y) = \nu(y)$, $\nabla\mu_t(y) = \nabla\nu(y)$. Then $\chi_1(\mu_t, y) = \chi_1(\nu, y)$, as they only involve derivatives up to order 1. Since $\|h(y)\| = \sqrt{\mu_t(y)} \geq \delta\epsilon_0$, then (13b) holds.

We proceed to show that the number of inner iterations is bounded by (12). Each outer iteration k of Algorithm 1 calls the inner method once. Let N_k be the number of inner iterations made in this call. The total number of inner iterations is $\sum_{k=1}^K N_k$.

We first analyze Phase I. The inner method is applied to the problem $\min_y \nu(y)$, starting with y_0 and terminating with y_1. By (11), we have

$$\nu(y_0) \;\geq\; \nu(y_0) - \nu(y_1) \;\geq\; N_1\, \kappa_\nu\, p(\boldsymbol{\epsilon}).$$

It follows that $N_1 \leq \nu(y_0)/\kappa_\nu p(\epsilon)$.

We proceed to Phase II. For each $a \in A$, let $n(a)$ be the next integer that lies in A. For the largest $a \in A$ we define $n(a) := K$, where K is the final iteration. We can group the indices $k \geq 2$ as follows:

$$\{2,3,\ldots,K\} = \bigcup_{a\in A} K_a, \qquad K_a := \{a{+}1, a{+}2, \ldots, n(a)\}.$$

We will show that for any $a \in A$ we have that

$$N(K_a) \;:=\; \sum_{k\in K_a} N_k \;\leq\; \epsilon_0^2/\kappa_\mu p(\epsilon). \tag{18}$$

Consider an iteration $k \in K_a$. The inner method is applied to $\min_y \mu(t_{k-1}, y)$, starting with y_{k-1} and terminating with y_k. By (11), we have

$$\mu(t_{k-1}, y_{k-1}) - \mu(t_{k-1}, y_k) \;\geq\; N_k\, \kappa_\mu\, p(\boldsymbol{\epsilon}).$$

Observe that $K_a \setminus \{n(a)\} \subset B$. By (16), we have

$$\mu(t_{k-1}, y_k) = \mu(t_k, y_k) \quad \text{ for } \quad k \in K_a \setminus \{n(a)\}.$$

Also note that $\mu(t_a, y_a) = \epsilon_0^2$ by (15). Therefore,

$$\begin{aligned}\epsilon_0^2 \;&\geq\; \mu(t_a, y_a) - \mu(t_{n(a)}{-}1, y_{n(a)})\\ &= \sum_{k\in K_a} \mu(t_{k-1}, y_{k-1}) - \mu(t_{k-1}, y_k) \geq \sum_{k\in K_a} N_k \kappa_\mu p(\boldsymbol{\epsilon}).\end{aligned}$$

By rearranging the above inequality we get (18).

Let us now upper bound the cardinality of A. By (15) and (16) we have that $t_{k-1}{-}t_k$ is at least $(1{-}\delta)\epsilon_0$ for $k \in A$, and is positive for $k \in B$. Also note that $t_0 = f(y_1)$ and $t_K \geq f(y_K){-}\epsilon_0 \geq f_{\text{low}}{-}\beta$ by (14). Then

$$f(y_1){-}f_{\text{low}}{+}\beta \;\geq\; t_0{-}t_K \;=\; \sum_{k=1}^{K}(t_{k-1}{-}t_k) \;\geq\; \sum_{k\in A}(t_{k-1}{-}t_k) \;\geq\; |A|\,(1{-}\delta)\epsilon_0,$$

and hence $|A| \leq (f(y_1){-}f_{\text{low}}{+}\beta)/(1{-}\delta)\epsilon_0$.

Combining everything, we derive

$$\sum_{k=1}^{K} N_k \;\leq\; N_1 + |A| \cdot \max_{a\in A} N(K_a) \;\leq\; \frac{\nu(y_0)}{\kappa_\nu p(\epsilon)} + \frac{f(y_1){-}f_{\text{low}}{+}\beta}{(1{-}\delta)\epsilon_0} \cdot \frac{\epsilon_0^2}{\kappa_\mu\, p(\epsilon)},$$

which is equal to (12).

A.3 Proof of Theorem 4

We finally show that AFAC points can be computed in polynomial time. Let $\varepsilon_0, \varepsilon_1, \varepsilon_2, \gamma, R_\lambda$ be as in the statement of Theorem 4. We consider Algorithm 1 with parameters

$$\begin{gathered}\delta \;:=\; 1/2, \qquad q \;:=\; 2, \qquad \boldsymbol{\epsilon} \;:=\; (\epsilon_1, \epsilon_2),\\ \epsilon_0 \;:=\; \varepsilon_0, \quad \epsilon_1 \;:=\; R_\lambda^{-1}\, \varepsilon_0\, \varepsilon_1, \quad \epsilon_2 \;:=\; \tfrac{1}{2} R_\lambda^{-1}\, \varepsilon_0\, \varepsilon_2.\end{gathered}$$

For the inner method we use the ARC algorithm from Theorem 3, using the criticality measure (10). Algorithm 1 returns a pair (y, t). The associated multiplier is $\lambda := (f(y){-}t)^{-1}\, h(y) \in \mathbb{R}^m$, which is defined only if $f(y) \neq t$.

In order to apply Theorem 8, we have to check that the functions ν and $\mu_t = \mu(t, \cdot)$ are smooth enough so that the inner algorithm satisfies (11).

Lemma 6 ([14, Lem.4.1]). *Assume that $\{\nabla^j f\}_{j=0}^q, \{\nabla^j h\}_{j=0}^q$ are uniformly bounded and Lipschitz continuous on a set $D \subset \mathbb{R}^n$. Then* *(i)* $\{\nabla^j \nu\}_{j=0}^q$ *are uniformly bounded and Lipschitz continuous on* D.

(ii) $\{\nabla^j \mu_t\}_{j=0}^q$ *are uniformly bounded and Lipschitz continuous on* $D \cap B_t$*, with* $B_t := \{y : |f(y)-t| \leq 1\}$*, and the constants are independent of* t.

The above lemma shows that ν is smooth on $\mathfrak{M}_\beta$ and μ_t is smooth on $\mathfrak{M}_\beta \cap B_t$, with $B_t := \{y : |f(y)-t| \leq 1\}$. Note that all points y_k produced by Algorithm 1 lie in $\mathfrak{M}_\beta \cap B_t$ because of (14). Since ν, μ_t are sufficiently smooth, we can apply Theorem 3 (see also [15]). We conclude that the inner method satisfies (11) with

$$p(\boldsymbol{\epsilon}) \;=\; \min\{\epsilon_1^2,\, \epsilon_2^3\} \;=\; \Omega(\min\{\varepsilon_0^2\,\varepsilon_1^2,\, \varepsilon_0^3\,\varepsilon_2^3\}).$$

Hence, by Theorem 8, the total number of inner iterations is $O(p(\boldsymbol{\epsilon})^{-1}) = O(\max\{\varepsilon_0^{-2}\,\varepsilon_1^{-2}, \varepsilon_0^{-3}\,\varepsilon_2^{-3}\})$. Since each inner iteration requires $O(1)$ function evaluations (see Theorem 3), then the total number of function evaluations has the same order of magnitude.

Let us see that the conditions (2) hold. Let (y,t) be the output of Algorithm 1. By Theorem 8, this pair satisfies either (13a) or (13b). Let us see that (13b) cannot occur. Assume that

$$\|h(y)\| > \epsilon_0/2, \qquad \|\nabla \nu(y)\| \leq \epsilon_1.$$

Observe that $\|h(y)\| \leq \epsilon_0 \leq \beta$ by (14), and hence ϱ-LICQ holds at y. Then

$$\varrho\,\epsilon_0 \;<\; 2\,\varrho\|h(y)\| \;\leq\; 2\,\|h(y)^T \nabla h(y)\| \;=\; \|\nabla \nu(y)\| \;\leq\; \epsilon_1.$$

Also note that

$$R_\lambda^{-1} \;\leq\; \tfrac{1}{2}\varrho(1{+}L_f)^{-1} \;\leq\; \tfrac{1}{2}\varrho,$$
$$\epsilon_1/\epsilon_0 \;=\; R_\lambda^{-1}\,\varepsilon_1 \;\leq\; \tfrac{1}{2}\varrho\,\varepsilon_1 \;\leq\; \tfrac{1}{2}\varrho.$$

The last two equations give a contradiction.

Then the output (y,t) satisfies (13a). Hence, $t < f(y)$ and

$$\|h(y)\| \;\leq\; \epsilon_0, \quad \|\nabla \mu_t(y)\| \;\leq\; \epsilon_1, \quad \nabla^2 \mu_t(y) \;\succeq\; -\epsilon_2 I_n.$$

Let $\alpha := (f(y){-}t)^{-1}$, so that $\lambda = \alpha\, h(y)$. It can be checked that $\alpha^2 \mu_t(y) = \|(1,\lambda)\|^2$. Note that $\mu_t(y) \geq (\epsilon_0/2)^2$ by (17), and hence

$$\alpha \;=\; \mu_t(y)^{-1/2}\,\|(1,\lambda)\| \;\leq\; 2\,\varepsilon_0^{-1}\|(1,\lambda)\|.$$

The Lagrangian function $L(y,\lambda) = f(y) + \lambda \cdot h(x)$ is closely related to $\mu_t(y)$. A simple calculation gives that

$$\begin{aligned} \nabla L(y,\lambda) \;&=\; \alpha \cdot \tfrac{1}{2}\nabla \mu_t(y), \\ \nabla^2 L(y,\lambda) \;&=\; \alpha\left(\tfrac{1}{2}\nabla^2 \mu_t(y) - \tilde{J}^T \tilde{J}\right), \end{aligned} \tag{19}$$

where $\tilde{J} := \begin{pmatrix} \nabla f(y) \\ \nabla h(y) \end{pmatrix}$ is the augmented Jacobian.

We proceed to verify (2a). We already have that $\|h(y)\| \leq \varepsilon_0$. Note that

$$\|\nabla L(y,\lambda)\| \;=\; \tfrac{1}{2}\,\alpha\,\|\nabla \mu_t(y)\| \;\leq\; \varepsilon_0^{-1}\|(1,\lambda)\|\,\epsilon_1 \;=\; R_\lambda^{-1}\|(1,\lambda)\|\,\varepsilon_1. \tag{20}$$

We claim that $\|(1,\lambda)\| \leq R_\lambda$. By Lemma 4 and $R_\lambda^{-1} \leq \varrho/2$, $\varepsilon_1 \leq 1$, we have

$$\|\lambda\| \;\leq\; \varrho^{-1}\,(R_\lambda^{-1}\,\varepsilon_1\|(1,\lambda)\| + \|\nabla f(y)\|) \;\leq\; \tfrac{1}{2}(1{+}\|\lambda\|) + \varrho^{-1}L_f.$$

It follows that $\|\lambda\| \leq 1 + 2\varrho^{-1}L_f = R_\lambda - 1$ and hence $\|(1,\lambda)\| \leq R_\lambda$, as we claimed. Then $\|\nabla L(y,\lambda)\| \leq \varepsilon_1$ by (20).

We now verify (2b). Let $u \in \mathbb{R}^n$ of unit norm such that $\|Ju\| \leq \gamma$, where $J := \nabla h(y)$. We need to show that $u^T \nabla^2 L(y,\lambda) u \geq -\varepsilon_2$. By (19), we have

$$u^T \nabla^2 L(y,\lambda) u \;=\; \alpha\,(\tfrac{1}{2}u^T \nabla^2 \mu_t(y) u - \|\tilde{J}u\|^2). \tag{21}$$

Note that $u^T \nabla^2 \mu_t(y) u \geq -\epsilon_2 = -\frac{1}{2}R_\lambda^{-1}\varepsilon_0\varepsilon_2$. We bound $\|\tilde{J}u\|$ next:

$$\tilde{J} \;=\; \begin{pmatrix} \nabla f(y) \\ \nabla h(y) \end{pmatrix} \;=\; \begin{pmatrix} \nabla L(y,\lambda) \\ 0 \end{pmatrix} + \begin{pmatrix} -\lambda^T J \\ J \end{pmatrix},$$

$$\|\tilde{J}u\| \;\leq\; \|\nabla L(y,\lambda)\| + \|(1,\lambda)\|\,\|Ju\| \;\leq\; \varepsilon_1 + \gamma\,\|(1,\lambda)\| \;\leq\; \tfrac{1}{2}(R_\lambda^{-1}\,\varepsilon_0\,\varepsilon_2)^{1/2},$$

where we used that ε_1 and γR_λ are at most $\frac{1}{4}(R_\lambda^{-1}\varepsilon_0\varepsilon_2)^{1/2}$ by (3). Hence

$$\alpha\,(\tfrac{1}{2}u^T \nabla^2 \mu_t(y) u - \|\tilde{J}u\|^2) \;\geq\; -(2\varepsilon_0^{-1}\|(1,\lambda)\|) \cdot (\tfrac{1}{2}R_\lambda^{-1}\varepsilon_0\,\varepsilon_2) \;\geq\; -\varepsilon_2.$$

Together with (21), we get that $u^T \nabla^2 L(y,\lambda) u \geq -\varepsilon_2$.

B Proofs from Section 4

Proof of Lemma 3. Let $L(X) := f(X) - \bar{S} \bullet X$, with $\bar{S} := S(\bar{X})$. This is a convex function with $\nabla L(\bar{X}) = 0$, so $\bar{X}$ is its global minimum. Note that

$$
\begin{aligned}
f(X) \;&=\; L(X) + \bar{S} \bullet X \;\geq\; L(X) - (\varepsilon_2 I_n) \bullet X \;\geq\; L(X) - \varepsilon_2 \|X\| \sqrt{n}, \\
L(\bar{X}) \;&=\; f(\bar{X}) - \bar{S} \bullet \bar{X} \;\geq\; f(\bar{X}) - \|\bar{S}\bar{X}\|_* \;\geq\; f(\bar{X}) - \varepsilon_1 \sqrt{n}.
\end{aligned}
$$

Since $L(X) \geq L(\bar{X})$, the result follows from the above equations. $\square$

The next lemma is an analogue of Lemma 2.

Lemma 7. *Let Y be an $(\varepsilon_1, \varepsilon_2)$-AC point of (BM$_{ls}$). If $\sigma_p(Y) \leq \sqrt{\varepsilon_2}/R_A$, then YY^T is $\boldsymbol{\varepsilon}'$-approximately optimal for (SDP$_{ls}$), with $\boldsymbol{\varepsilon}' := (0, R_Y \varepsilon_1, 5\varepsilon_2)$.*

Proof. Let Y satisfy (9), and let us show that YY^T satisfies (8). The first-order condition is easy to check. We proceed to show that $u^T S(X) u \geq -\varepsilon_2'$ for any unit vector $u \in \mathbb{R}^n$. Let $z \in \mathbb{R}^p$ be a unit vector such that $\|Yz\| = \sigma_p(Y)$. The matrix $U := uz^T$ satisfies $\|U\| = 1$ and $\|UY^T\| \leq \|u\| \|Yz\| = \sigma_p(Y)$. Then $\|\mathcal{A}(UY^T)\| \leq \sqrt{\varepsilon_2}$ and by (9b) we have

$$u^T S(X) u = S(YY^T) \bullet UU^T \geq -\varepsilon_2 - 4\|\mathcal{A}(UY^T)\|^2 \geq -5\varepsilon_2. \qquad \square$$

Proof of Proposition 2. As $\mathcal{A} \in \mathscr{A}_{\boldsymbol{\varepsilon}}$, there is a spurious $\boldsymbol{\varepsilon}$-AC point Y. By Lemma 7, we must have $\sigma_p(Y) > \sqrt{\varepsilon_2}/R_A$. Note that $\|S(YY^T)Y\| \leq \varepsilon_1$. Together with (6), we conclude that $S(YY^T) \in \operatorname{tube}_\delta(\mathbb{S}^n_{n-p})$. Let $\lambda := 2(\mathcal{A}(YY^T) - b)$. Note that $\|\lambda\| > 2\varepsilon_0$ and

$$\|\lambda\| = 2\, \|\mathcal{A}(YY^T) - b\| \leq 2(\|\mathcal{A}\| \|Y\|^2 + \|b\|) \leq R_\lambda.$$

Then $\lambda \in D_\lambda$ and $\mathcal{A}^*(\lambda) = S(YY^T) \in \operatorname{tube}_\delta(\mathbb{S}^n_{n-p})$. $\square$

Proof of Theorem 7. The result in Proposition 2 can be expressed as:

$$\mathscr{A}_{\boldsymbol{\varepsilon}} \;\subset\; \{\mathcal{A} \in (\mathbb{S}^n)^m : 0 \in \operatorname{tube}_\delta(\mathbb{S}^n_{n-p}) + \mathcal{A}^*(D_\lambda)\},$$

which is closer to the formula in Proposition 1. Consider an ϵ-net $\mathcal{N}$ of D_λ, where $\epsilon := \delta/R_A$. It suffices to take $(3R_\lambda/\epsilon)^m = (3\kappa/\delta)^m$ points for the ϵ-net. A reasoning similar to (7) gives

$$
\begin{aligned}
\mathscr{A}_{\boldsymbol{\varepsilon}} \;&\subset\; \{\mathcal{A} \in (\mathbb{S}^n)^m : 0 \in \operatorname{tube}_{2\delta}(\mathbb{S}^n_{n-p}) + \mathcal{A}^*(\mathcal{N}))\} \\
&=\; \bigcup_{\ell \in \mathcal{N}} \{\mathcal{A} \in (\mathbb{S}^n)^m : \mathcal{A}^*(\ell) \in \operatorname{tube}_{2\delta}(\mathbb{S}^n_{n-p}))\}.
\end{aligned}
$$

Let $\ell \in \mathcal{N}$, and consider the linear map

$$\phi_\ell : (\mathbb{S}^n)^m \to \mathbb{S}^n, \quad \mathcal{A} \mapsto \mathcal{A}^*(\ell).$$

This is a surjective map. Moreover, the scaled map $\frac{1}{\|\ell\|}\, \phi_\ell$ gives an isometry $(\ker \phi_\ell)^\perp \cong \mathbb{S}^n$. It follows that

$$\phi_\ell(\mathcal{A}) \;\in\; \operatorname{tube}_{2\delta}(\mathbb{S}^n_{n-p}) \quad\iff\quad \mathcal{A} \;\in\; \operatorname{tube}_{2\delta/\|\ell\|}\left(\phi_\ell^{-1}(\mathbb{S}^n_{n-p})\right).$$

Since $\|\ell\| \geq 2\varepsilon_0$, we conclude that

$$\mathscr{A}_{\boldsymbol{\varepsilon}} \;\subset\; \bigcup_{\ell \in \mathcal{N}} \operatorname{tube}_{\delta/\varepsilon_0}(V_\ell), \quad \text{with } \; V_\ell := \phi_\ell^{-1}(\mathbb{S}^n_{n-p}).$$

The final part of the proof is similar to the one in Theorem 5. The variety V_ℓ is a cylinder over $\mathbb{S}^n_{n-p}$, so it has the same codimension $\tau(p)$ and degree $n-p+1$ as $\mathbb{S}^n_{n-p}$, The ambient space is $(\mathbb{S}^n)^m$, of dimension $\tau(n)m$. Using the union bound and Theorem 6, we get

$$\Pr[\mathcal{A} \in \mathscr{A}_{\boldsymbol{\varepsilon}}] \;<\; \#\mathcal{N} \cdot \Pr[\mathcal{A} \in \operatorname{tube}_{\delta/\varepsilon_0}(V_\ell)] \;<\; (3\kappa/\delta)^m \cdot 4e \left(2n^3 m\delta/\sigma\varepsilon_0\right)^{\tau(p)}. \qquad \square$$

C Explicit complexity estimates

In this section we provide explicit complexity estimates for Theorems 1 and 2. We first introduce some notation. Consider constants $\alpha \geq \beta > 0$ and sets $\mathfrak{M}_\alpha \supset \mathfrak{M}_\beta$, where $\mathfrak{M}_t := \{Y : \|\mathcal{A}(YY^T){-}b\| \leq t\}$. We assume that β is small enough so that ϱ-LICQ holds globally on $\mathfrak{M}_\beta$. On the other hand, $\alpha > 0$ is sufficiently large so that a point $Y_0 \in \mathfrak{M}_\alpha$ is always known. We further assume that $\mathfrak{M}_\alpha$ is compact. This is satisfied, for instance, when the feasible set of (*SDP*) is compact and satisfies Slater's condition, as shown next.

Lemma 8. *Assume that the set $\{X : \mathcal{A}(X){=}b, X{\succeq}0\}$ is compact and satisfies Slater's condition (i.e., $\exists X : \mathcal{A}(X){=}b, X{\succ}0$). Then $\mathfrak{M}_t$ is compact for any $t \geq 0$.*

Proof. Consider the SDP $\max\{I \bullet X : \mathcal{A}(X){=}b, X{\succeq}0\}$ and its dual $\min\{b^T\lambda : \mathcal{A}^*(\lambda){\succeq}I\}$. Let R_0^2 be the primal optimal value, which is finite by compactness. Strong duality holds by Slater's condition. So the dual optimum is attained at some $\bar\lambda$, and $\mathcal{A}^*(\bar\lambda) \succeq I$, $b^T\bar\lambda = R_0^2$. Given $Y \in \mathcal{M}_t$,

$$\|Y\|^2 = I \bullet YY^T \leq \mathcal{A}^*(\bar\lambda) \bullet YY^T = \bar\lambda \cdot \mathcal{A}(YY^T) = \bar\lambda \cdot b + \bar\lambda \cdot (\mathcal{A}(YY^T){-}b) \leq R_0^2 + t\|\bar\lambda\|.$$

We conclude that $\mathcal{M}_t$ is contained in a ball of radius $(R_0^2 + t\|\bar\lambda\|)^{1/2}$. $\square$

Notice that a suitable value α can be obtained from an arbitrary point $Y_0 \in \mathbb{R}^{n\times p}$. On the other hand, β should be $\Omega(\varrho_{\min})$, where $\varrho_{\min}$ is the smallest LICQ constant among all feasible points $Y \in \mathfrak{M}_0$.

C.1 Solving (*SDP*)

Assume that an approximately feasible solution Y_0 is known. Consider the following setting:

- p satisfies $\tau(p) \geq (1{+}\eta)m{+}\eta\, t$ for some given constants $\eta, t{\in}\mathbb{R}_+$.
- $\mathcal{A}, b$ are fixed and C is uniformly distributed on a ball $\mathbf{B}_\sigma(\bar C)$.
- $\exists \beta \in \mathbb{R}_+$ such that: $\mathfrak{M}_\beta$ is compact, a point $Y_0 \in \mathfrak{M}_\beta$ is known, and ϱ-LICQ holds on $\mathfrak{M}_\beta$.
- $R_Y, L_f \in \mathbb{R}_+$ are constants that bound $\|Y\|$ and $\|CY\|$, for $Y \in \mathfrak{M}_\beta$.
- (*BM*) is solved with the method from Theorem 4 initialized at Y_0.

The next theorem shows that the Burer-Monteiro method solves (*SDP*) in polynomial time with high probability.

Theorem 9. *Let $\rho{\in}(0, 1]$ arbitrary, and let*

$$\varepsilon_0 := \gamma := \epsilon, \quad \varepsilon_1 := \epsilon^2, \quad \varepsilon_2 := 16\, R_\lambda^3\, \epsilon,$$
$$\text{with} \quad \epsilon := K^{-1}\,\rho\,(\sigma/4n^3)^{1+1/\eta},$$

where R_λ and K are the problem dependent constants

$$R_\lambda := 2 + 2\varrho^{-1}L_f, \quad K := \|\mathcal{A}\|\,(3\kappa)^{1/\eta}, \quad \kappa := R_\lambda\|\mathcal{A}\|.$$

The algorithm from Theorem 4 returns a pair (Y, λ) after $O(\epsilon^{-6})$ function evaluations. With probability at least $1{-}O(\sigma/n^3)^t$, the pair (YY^T, λ) is $(\epsilon, \epsilon^2 R_Y, 16R_\lambda^3\epsilon)$-approximately optimal for (SDP).

Proof. The smoothness assumptions in Theorem 4 are satisfied since $\mathfrak{M}_\beta$ is compact. Then (Y, λ) is an $(\boldsymbol{\varepsilon}, \gamma)$-AFAC pair with $\|\lambda\|{\leq}R_\lambda$. Note that

$$\delta \;:=\; \varepsilon_1\|\mathcal{A}\|/\gamma \;=\; \epsilon\,\|\mathcal{A}\| \;\leq\; (1/3\kappa)^{1/\eta}\,(\sigma/2en^3)^{1+1/\eta}$$

is as in Corollary 1. Hence (YY^T, λ) is $(\varepsilon_0, \varepsilon_1 R_Y, \varepsilon_2)$-approximately optimal for (*SDP*) with probability $1{-}O(\sigma/n^3)^t$. $\square$

The above theorem shows that YY^T obtained is approximately optimal for the perturbed problem (*SDP*) with high probability. Let ($\overline{SDP}$) denote the SDP problem in which we use the unperturbed cost matrix $\bar C$. We can also show that YY^T is also approximately optimal for ($\overline{SDP}$).

Corollary 3. *Consider the setup of Theorem 9. With probability at least $1{-}O(\sigma/n^3)^t$, the pair (YY^T, λ) is $(\varepsilon_0'', \varepsilon_1'', \varepsilon_2'')$-approximately optimal for ($\overline{SDP}$), where $\varepsilon_0'', \varepsilon_1'', \varepsilon_2'' = O(\sigma)$.*

Proof. Let $X := YY^T$. We know that (X, λ) is $(\varepsilon_0', \varepsilon_1', \varepsilon_2')$-approximately optimal for (*SDP*) with high probability. Let $S := C - \mathcal{A}^*(\lambda)$, $\bar{S} := \bar{C} - \mathcal{A}^*(\lambda)$ be the slack matrices for (*SDP*) and $(\overline{SDP})$. Observe that

$$\|\mathcal{A}(X) - b\| \le \varepsilon_0' \le O(\sigma), \tag{22}$$

$$\|\bar{S}X\| \le \|SX\| + \|(\bar{S} - S)X\| \le \varepsilon_1' + \sigma\|X\| \le O(\sigma), \tag{23}$$

$$\bar{S} \succeq S - \|\bar{S} - S\| I_n \succeq -(\varepsilon_2' + \sigma) I_n \succeq -O(\sigma) I_n. \tag{24}$$

So the optimality conditions of $(\overline{SDP})$ hold with $\varepsilon_0'', \varepsilon_1'', \varepsilon_2'' = O(\sigma)$. □

C.2 Solving (SDP_{ls})

Consider the following setting:

- p satisfies $\tau(p) \ge (1+\eta)m + \eta\, t$ for some given constants $\eta, t \in \mathbb{R}_+$.
- b is fixed and $\mathcal{A}$ is uniformly distributed on a ball $\mathbf{B}_\sigma(\bar{\mathcal{A}})$.
- $\exists \alpha \in \mathbb{R}_+$ and a matrix Y_0 such that $\mathfrak{M}_\alpha$ is compact and $Y_0 \in \mathfrak{M}_\alpha$
- $R_Y \in \mathbb{R}_+$ is a constant that bounds $\|Y\|$, for $Y \in \mathfrak{M}_\alpha$.
- (BM_{ls}) is solved with the method from Theorem 3 initialized at Y_0.

Theorem 10. *Let $\rho \in (0, 1]$ arbitrary, and let*

$$\varepsilon_1 := \epsilon^{3/2}, \quad \varepsilon_2 := \epsilon, \quad \epsilon := K^{-1} \left(\rho\, \sigma^2 / 2n^3 m\right)^{1+1/\eta},$$

where $K := R_A\, (3\kappa)^{1/\eta}$, expressed in terms of

$$\kappa := 2(R_A R_Y^2 + \|b\|) R_A, \quad R_A := \|\bar{\mathcal{A}}\| + \sigma.$$

The algorithm from Theorem 3 returns a point Y after $O(\epsilon^{-3})$ function evaluations. With probability at least $1 - O(\sigma^2/n^3 m)^t$, we have that YY^T is $(\rho\,\sigma, \epsilon^{3/2} R_Y, 5\epsilon)$-approximately optimal for (SDP_{ls}).

Proof. The smoothness assumptions in Theorem 3 are satisfied since $\mathfrak{M}_\alpha$ is compact. Therefore Y is an $(\varepsilon_1, \varepsilon_2)$-AC point. Note that

$$\delta := \varepsilon_1 R_A / \sqrt{\varepsilon_2} = \epsilon\, R_A = (1/3\kappa)^{1/\eta} (\rho\, \sigma^2 / 2n^3 m)^{1+1/\eta}$$

is as in Corollary 2. Hence YY^T is $(\rho\,\sigma, \varepsilon_1 R_Y, 5\varepsilon_2)$-approximately optimal for (SDP_{ls}) with probability $1 - O(\sigma^2/n^3 m)^t$. □

Remark. The above theorem holds even if the optimal value of (SDP_{ls}) is nonzero. In the special case that the optimal value is zero, then by Lemma 3 we have that

$$\|\mathcal{A}(YY^T) - b\| \le \max\{\varepsilon_0',\ n^{1/4}(\varepsilon_1' + \varepsilon_2' R_Y)^{1/2}\},$$

where $\varepsilon_0' = \rho\,\sigma$, $\varepsilon_1' = \epsilon^{3/2} R_Y$, $\varepsilon_2' = 5\epsilon$ are the optimality constants from Theorem 10.

Let $(\overline{SDP}_{ls})$ denote the instance of problem (SDP_{ls}) in which we use the unperturbed constraints $\bar{\mathcal{A}}$. We next show that YY^T is also approximately optimal for $(\overline{SDP}_{ls})$.

Corollary 4. *Consider the setup of Theorem 10. With probability at least $1 - O(\sigma^2/n^3 m)^t$, the matrix YY^T is $(\varepsilon_0'', \varepsilon_1'', \varepsilon_2'')$-approximately optimal for $(\overline{SDP}_{ls})$, where $\varepsilon_0'', \varepsilon_1'', \varepsilon_2'' = O(\sigma)$.*

Proof. We know that the matrix $X := YY^T$ is $(\varepsilon_0', \varepsilon_1', \varepsilon_2')$-approximately optimal for (*SDP*) with high probability. There are two cases. The first case is that $\|\mathcal{A}(X) - b\| \le \varepsilon_0'$, which implies that

$$\|\bar{\mathcal{A}}(X) - b\| \le \|\mathcal{A}(X) - b\| + \|(\bar{\mathcal{A}} - \mathcal{A})X\| \le \varepsilon_0' + \sigma\|X\| \le O(\sigma).$$

This means that $\varepsilon_0'' = O(\sigma)$. Consider the variables:

$$\lambda := 2(\mathcal{A}(X) - b), \quad S := \mathcal{A}^*(\lambda),$$

$$\bar{\lambda} := 2(\bar{\mathcal{A}}(X) - b), \quad \bar{S} := \bar{\mathcal{A}}^*(\bar{\lambda}).$$

The second case is that $\|SX\| \le \varepsilon_1$, $S \succeq -\varepsilon_2 I_n$. Note that

$$\|\bar{\lambda} - \lambda\| \le 2\, \|\bar{\mathcal{A}} - \mathcal{A}\|\, \|X\| \le O(\sigma),$$

$$\|\bar{S} - S\| \le \|\bar{\mathcal{A}}^*(\bar{\lambda} - \lambda)\| + \|(\bar{\mathcal{A}}^* - \mathcal{A}^*)\lambda\| \le O(\sigma).$$

From (23) and (24) we get that $\|\bar{S}X\| \le O(\sigma)$ and $\bar{S} \succeq -O(\sigma) I_n$. So the optimality conditions of $(\overline{SDP}_{ls})$ hold with $\varepsilon_0'', \varepsilon_1'', \varepsilon_2'' = O(\sigma)$. □