March 26, 2018

Lecture 11 – Advanced topics

A Induced Dimension Reduction

Let $\mathbf{A}$ be a $n \times n$ matrix and let $\mathbf{R}$ be a full rank $n \times s$ matrix: $\mathbf{R}$ is the so-called **IDR test matrix**. $\mathbf{R}^\perp$ is the subspace of all vectors $\mathbf{v}$ that are orthogonal to all column vectors of $\mathbf{R}$.

The following 'IDR theorem' indicates a way to 'reduce' the dimension of a subspace.

Theorem 11.1 *Let (ω_k) be a sequence of non-zero scalars. With $\mathcal{G}_0 \equiv \mathbb{C}^n$ (or a subspace of $\mathbb{C}^n$ that is invariant under multiplication by $\mathbf{A}$), let the sequence $(\mathcal{G}_k)$ of subspace be defined by*

$$\mathcal{G}'_k \equiv \mathcal{G}_k \cap \mathbf{R}^\perp, \quad \mathcal{G}_{k+1} \equiv (\mathbf{I} - \omega_k \mathbf{A})\mathcal{G}'_k. \tag{11.1}$$

Then we have that

$$\mathcal{G}_{k+1} \subset \mathcal{G}_k, \qquad \text{and} \qquad \mathbf{A}\mathcal{G}'_k \subset \mathcal{G}_k. \tag{11.2}$$

If the subspace $\mathbf{R}^\perp$ does not contain an eigenvector of $\mathbf{A}$, then we have

$$\mathcal{G}_{k+1} = \mathcal{G}_k \quad \Leftrightarrow \quad \mathcal{G}_k = \{\mathbf{0}\}. \tag{11.3}$$

By selecting $\mathbf{R}$ randomly, the probability that $\mathbf{R}^\perp$ contains an eigenvector is 0.

The formulae in (11.1) define a chain of IDR subspaces with strictly decreasing dimension (according to (11.2) and (11.3)). In particular $\mathcal{G}_k = \{\mathbf{0}\}$ for k at most n. **IDR methods** (induced dimension reduction) iteratively solve linear equations $\mathbf{Ax} = \mathbf{b}$ by constructing residuals $\mathbf{r}_k$ in $\mathcal{G}_k$. Then $\mathbf{r}_k = \mathbf{0}$ for some $k \leq n$. In practice $\|\mathbf{r}_k\|_2 \ll \|\mathbf{r}_0\|_2$ often already for modest values of k ($k \ll n$).

In the basic IDR method, ω_k is selected as in Bi-CGSTAB to minimise the residual norm $\omega_k = \text{argmin}_\omega \|\mathbf{r}'_k - \omega \mathbf{A}\mathbf{r}'_k\|_2$. For $s = 1$, this method is mathematically (as well as computationally) equivalent to Bi-CGSTAB. Actually IDR(s) (for $s > 1$) can be viewed as a Bi-CGSTAB version where the initial shadow residual $\widetilde{\mathbf{r}}_0$ is replaced by an $n \times s$ matrix $\mathbf{R}$, that is a one dimensional space by an s-dimensional one (the vectors are not of importance, but space that they span).

Exercise 11.1. ***Proof of Theorem*** 11.1**.** Let $\mu_0 \in \mathbb{C}$. For a subspace $\mathcal{S}_0$ of $\mathbb{C}^n$, put

$$\mathcal{S}'_0 \equiv \mathcal{S}_0 \cap \mathbf{R}^\perp \equiv \{\mathbf{v} \in \mathcal{S}_0 \,\big|\, \mathbf{v} \perp \mathbf{R}\}, \quad \mathcal{S}_1 \equiv (\mathbf{A} - \mu_0 \mathbf{I})\mathcal{S}'_0.$$

(a) Show that $\mathcal{S}_1 \subset \mathcal{S}_0$ implies that $(\mathbf{A} - \mu_1 \mathbf{I})(\mathcal{S}_1 \cap \mathbf{R}^\perp) \subset \mathcal{S}_1$ for all $\mu_1 \in \mathbb{C}$.

(b) Prove that $\mathcal{S}'_0$ contains an eigenvector of $\mathbf{A}$ if $\mathcal{S}_1 = \mathcal{S}_0 \neq \{\mathbf{0}\}$.
(Hint: Prove that $\dim(\mathcal{S}_0) = \dim(\mathcal{S}_0 \cap \mathbf{R}^\perp)$.)

(c) Prove that $\mathcal{G}_{k+1} \subset \mathcal{G}_k$ and $\mathcal{G}_{k+1} = \mathcal{G}_k \quad \Leftrightarrow \quad \mathcal{G}_k = \{\mathbf{0}\}$.

(d) Prove that $\mathbf{A}\mathcal{G}'_k \subset \mathcal{G}_k$.

The IDR spaces are related to Krylov and block Krylov subspaces. This relation explains how IDR methods and Bi-CGSTAB are related.

Consider the **block Krylov subspace** generated by $\mathbf{A}^*$ and $\mathbf{R}$

$$\mathcal{K}_k(\mathbf{A}^*, \mathbf{R}) \equiv \left\{ \sum_{j=0}^{k-1} (\mathbf{A}^*)^j \, \mathbf{R} \, \vec{\gamma}_j \,\Big|\, \vec{\gamma}_j \in \mathbb{C}^k \right\}.$$

Note that in standard Krylov subspace $\mathbf{R}$ is 1-dimensional.

For a polynomial q we define the **Sonneveld subspace**

$$\mathcal{S}(q, \mathbf{A}, \mathbf{R}) \equiv \{q(\mathbf{A})\mathbf{v} \,\big|\, \mathbf{v} \in \mathcal{G}_0, \ \mathbf{v} \perp \mathcal{K}_k(\mathbf{A}^*, \mathbf{R})\}, \quad \text{where } k \text{ is the degree of } q.$$

The IDR spaces of Theorem 11.1 are Sonneveld subspaces.

Theorem 11.2 *With* (ω_k) *and* $(\mathcal{G}_k)$ *as in Theorem* 11.1 *and polynomials* p_k *defined by*

$$p_0(\zeta) \equiv 1, \quad p_k(\zeta) \equiv (1 - \omega_{k-1}\zeta)\, p_{k-1}(\zeta) \qquad (\zeta \in \mathbb{C}, k = 0, 1, 2, \ldots)$$

we have that

$$\mathcal{G}_k = \mathcal{S}(p_k, \mathbf{A}, \mathbf{R}). \tag{11.4}$$

Exercise 11.2. Prove Theorem 11.2.

As mentioned, IDR methods construct the residual $\mathbf{r}_k$ in $\mathcal{G}_k$. In this context, the polynomial p_k of (11.4) is called the **stabilisation polynomial**.

The following exercise explains how Sonneveld subspaces are related to block rational Krylov subspaces. This insight may help to understand the excellent convergence properties of IDR methods.

Exercise 11.3. **Sonneveld subspaces and rational Krylov subspaces.** Let (ω_j), $(\mathcal{G}_k)$ and p_k be as defined in Theorem 11.2. Assume that none of the $1/\omega_j$ is an eigenvalue of $\mathbf{A}$. Consider the block **rational Krylov subspace**

$$\mathcal{K}_k(\mathbf{A}^*, \bar{p}_k(\mathbf{A}^*)^{-1}\mathbf{R}). \tag{11.5}$$

(a) Show that this is the orthogonal complement of the space in (11.4):

$$\mathcal{G}_k = (\mathcal{K}_k(\mathbf{A}^*, \bar{p}_k(\mathbf{A}^*)^{-1}\mathbf{R}))^{\perp}.$$

(b) Prove that

$$\mathcal{K}_k(\mathbf{A}^*, (\bar{p}_k(\mathbf{A}^*))^{-1}\mathbf{R}) = \text{span}\,((\mathbf{I} - \bar{\omega}_0\mathbf{A})^{-1}\mathbf{R}, \ldots, (\bar{p}_{k-1}(\mathbf{A}^*))^{-1}\mathbf{R}, (\bar{p}_k(\mathbf{A}^*))^{-1}\mathbf{R}).$$

(c) Assume that, in addition, the ω_j are mutually different. Show that

$$\mathcal{K}_k(\mathbf{A}^*, (\bar{p}_k(\mathbf{A}^*))^{-1}\mathbf{R}) = \text{span}\,((\mathbf{I} - \bar{\omega}_0\mathbf{A}^*)^{-1}\mathbf{R}, \ldots, (\mathbf{I} - \bar{\omega}_{k-2}\mathbf{A}^*)^{-1}\mathbf{R}, (\mathbf{I} - \bar{\omega}_{k-1}\mathbf{A}^*)^{-1}\mathbf{R}).$$

(d) Prove that $\mathcal{K}_{k+1}(\mathbf{A}^*, \mathbf{R}) = \text{span}\,(\mathbf{R}, \bar{p}_1(\mathbf{A}^*)\mathbf{R}, \ldots, \bar{p}_{k-1}(\mathbf{A}^*)\mathbf{R}, \bar{p}_k(\mathbf{A}^*)\mathbf{R})$. The Krylov subspace does not depend on the shifts. Does a similar statement hold for rational Krylov subspaces?

Basically, IDR methods proceed as follows.
Assume the vectors $\mathbf{r}_k, \mathbf{x}_k$ and $n \times s$ matrices $\mathbf{c}_k, \mathbf{u}_k$ are available with

$$\mathbf{r}_k, \mathbf{c}_k \in \mathcal{G}_k \quad \text{and} \quad \mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k, \quad \mathbf{c}_k = \mathbf{A}\mathbf{u}_k.$$

Here, with $\mathbf{c}_k \in \mathcal{G}_k$, we mean that the columns of $\mathbf{c}_k$ are vectors in $\mathcal{G}_k$. To construct a residual $\mathbf{r}_{k+1}$ and a matrix $\mathbf{c}_{k+1}$ in $\mathcal{G}_{k+1}$, consider the skew projections

$$\Pi_1 \equiv \mathbf{I} - \mathbf{c}_k\sigma^{-1}\mathbf{R}^* \quad \text{with} \quad \sigma \equiv \mathbf{R}^*\mathbf{c}_k, \qquad \Pi_0 = \mathbf{I} - \mathbf{u}_k\sigma^{-1}\mathbf{R}^*\mathbf{A}. \tag{11.6}$$

Π_1 projects along $\text{span}(\mathbf{c}_k)$ orthogonal to $\mathbf{R}$. Then, following the definitions in (11.1), we have

$$\mathbf{r}'_k \equiv \Pi_1\mathbf{r}_k = \mathbf{r}_k - \mathbf{c}_k\sigma^{-1}\mathbf{R}^*\mathbf{r}_k \in \mathcal{G}'_k \quad \text{and} \quad \mathbf{r}_{k+1} \equiv (\mathbf{I} - \omega_k\mathbf{A})\mathbf{r}'_k \in \mathcal{G}_{k+1}. \tag{11.7}$$

Updating $\mathbf{x}_k$ is easy. We simply update $\widetilde{\mathbf{x}}$ by $+\mathbf{u}$ if we update $\widetilde{\mathbf{r}}$ by $-\mathbf{c}$ and $\mathbf{c} = \mathbf{A}\mathbf{u}$:

$$\mathbf{x}'_k \equiv \mathbf{x}_k + \mathbf{u}_k\sigma^{-1}\mathbf{R}^*\mathbf{r}_k \quad \text{and} \quad \mathbf{x}_{k+1} \equiv \mathbf{x}'_k + \omega_k\mathbf{r}'_k.$$

Then, we have that $\mathbf{r}_{k+1} = \mathbf{b} - \mathbf{A}\mathbf{x}_{k+1}$. Note that $\sigma^{-1}(\mathbf{R}^*\mathbf{r}_k)$ has to be computed only once.

Since $\Pi_1 \mathbf{c}_k = \mathbf{0}$ (why?) we can not use the 'lifting technique' in(11.7) for constructing $\mathbf{r}_{k+1}$ to obtain $\mathbf{c}_{k+1}$. However, for the columns of $\mathbf{c}'_k$, we can take a(ny) basis of $\Pi_1\mathbf{A}\mathcal{K}_s(\Pi_1\mathbf{A}, \mathbf{r}'_k)$ (for details, see Exercise 11.5). Then $\mathbf{c}_{k+1} = (\mathbf{I} - \omega_k\mathbf{A})\mathbf{c}'_k$. Since

$$\Pi_1\mathbf{A} = \Pi_0\mathbf{A}$$

we have that $\Pi_1\mathbf{A}\mathcal{K}_s(\Pi_1\mathbf{A}, \mathbf{r}'_k) = \mathbf{A}\Pi_0\mathcal{K}_s(\Pi_1\mathbf{A}, \mathbf{r}'_k)$ which shows that there is a $\mathbf{u}'_k$ such that $\mathbf{c}'_k = \mathbf{A}\mathbf{u}'_k$: the columns of $\mathbf{u}'_k$ span $\Pi_0\mathcal{K}_s(\Pi_1\mathbf{A}, \mathbf{r}'_k)$.

There is a lot of freedom in constructing a basis of $\Pi_1\mathbf{A}\mathcal{K}_s(\Pi_1\mathbf{A}, \mathbf{r}'_k)$. This freedom can be exploited to improve efficiency and to enhance stability. The original IDR(s) method uses the residual updates that appear when using s-steps of Richardson (with parameter $\alpha = \omega_k$) for solving $\Pi_1\mathbf{A}\mathbf{z} = \mathbf{r}'_k$ with $\mathbf{z}_0 = \mathbf{x}_k$.

The initial matrix $\mathbf{u}_0$ (and $\mathbf{c}_0 = \mathbf{A}\mathbf{u}_0$) can be generated in a similar way, with columns spanning $\mathcal{K}_s(\mathbf{A}, \mathbf{r}_0)$. With this choice, the residuals $\mathbf{r}_k$ will be in the Krylov subspace $\mathcal{K}_{k(s+1)}(\mathbf{A}, \mathbf{r}_0)$. Although the approach sketched above allows $\mathbf{u}_0$ to be any $n \times s$ matrix, the Krylov subspace start seems to lead to the best convergence results.

The following exercises discusses the details for the above approach in case the IDR test matrix $\mathbf{R} = [\widetilde{\mathbf{r}}_0]$ is 1-dimensional. In the subsequent exercise, we consider a general IDR test matrix $\mathbf{R}$

Exercise 11.4. **IDR methods.** In this exercise, we assume that $\mathbf{R} = [\,\widetilde{\mathbf{r}}_0\,]$, i.e., $s = 1$. We construct residuals $\mathbf{r}_k$ in $\mathcal{G}_k$ for the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$.

Assume the vectors $\mathbf{r}_k, \mathbf{x}_k, \mathbf{c}_k, \mathbf{u}_k$ are available with

$$\mathbf{r}_k, \mathbf{c}_k \in \mathcal{G}_k \quad \text{and} \quad \mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k, \quad \mathbf{c}_k = \mathbf{A}\mathbf{u}_k.$$

We construct a residual $\mathbf{r}_{k+1}$ and a vector $\mathbf{c}_{k+1}$ in $\mathcal{G}_{k+1}$. Let Π_1 and Π_0 be defined as in (11.6). Note that now $\sigma \equiv \widetilde{\mathbf{r}}_0^*\mathbf{c}_k$.

(a) Put

$$\mathbf{r}'_k \equiv \mathbf{r}_k - \mathbf{c}_k\sigma^{-1}\rho, \quad \text{where} \quad \rho \equiv \mathbf{R}^*\mathbf{r}_k = \widetilde{\mathbf{r}}_0^*\mathbf{r}_k \quad \text{and} \quad \mathbf{r}_{k+1} \equiv \mathbf{r}'_k - \omega_k\mathbf{A}\mathbf{r}'_k.$$

Prove that

i) $\mathbf{r}'_k = \Pi_1\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}'_k \in \mathcal{G}'_k$, where $\mathbf{x}'_k \equiv \mathbf{x}_k + \mathbf{u}_k\sigma^{-1}\rho$,

ii) $\mathbf{r}_{k+1} = (\mathbf{I} - \omega_k\mathbf{A})\mathbf{r}'_k = \mathbf{b} - \mathbf{A}\mathbf{x}_{k+1} \in \mathcal{G}_{k+1}$, where $\mathbf{x}_{k+1} \equiv \mathbf{x}'_k + \omega_k\mathbf{r}'_k$

(b) Why can we not construct a $\mathbf{c}'_k$ in $\mathcal{G}'_k$ as $\mathbf{c}'_k = \Pi_1\mathbf{c}_k$, analogue to the construction of $\mathbf{r}'_k$?

(c) Show that $\mathbf{v} \equiv \mathbf{A}\mathbf{r}'_k \in \mathcal{G}_k$. Let

$$\mathbf{c}'_k \equiv \mathbf{v} - \mathbf{c}_k\sigma^{-1}\mu \quad \text{with} \quad \mu \equiv \mathbf{R}^*\mathbf{v} = \widetilde{\mathbf{r}}_0^*\mathbf{v}, \qquad \mathbf{c}_{k+1} \equiv \mathbf{c}'_k - \omega_k\mathbf{A}\mathbf{c}'_k.$$

Show that

i) $\mathbf{c}'_k = \Pi_1\mathbf{v} = \mathbf{A}\mathbf{u}'_k \in \mathcal{G}'_k$, where $\mathbf{u}'_k \equiv \mathbf{r}'_k - \mathbf{u}_k\sigma^{-1}\mu = \Pi_0\mathbf{r}'_k$

ii) $\mathbf{c}_{k+1} = \mathbf{A}\mathbf{u}_{k+1} \in \mathcal{G}_{k+1}$, where $\mathbf{u}_{k+1} \equiv \mathbf{u}'_k - \omega_k\mathbf{c}'_k$.

Note that $\sigma^{-1}\mu$ can be used for the computation of $\mathbf{c}'_k$ as well as $\mathbf{u}'_k$.

(d) Now take $\omega_k \equiv \text{argmin}_\omega\|\mathbf{r}'_k - \omega\mathbf{A}\mathbf{r}'_k\|_2$. Give an (efficient) algorithm for the iterative process indicated above. How do you initiate the iteration? Compare this algorithm with Bi-CGSTAB (do the comparison theoretically as well as experimentally).

Exercise 11.5. **IDR methods II.** We continue Exercise 11.4, now with a general $n \times s$ matrix $\mathbf{R}$.

Assume the vectors $\mathbf{r}_k, \mathbf{x}_k$ and $n \times s$ matrices $\mathbf{c}_k, \mathbf{u}_k$ are available with

$$\mathbf{r}_k, \mathbf{c}_k \in \mathcal{G}_k \quad \text{and} \quad \mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k, \quad \mathbf{c}_k = \mathbf{A}\mathbf{u}_k.$$

We construct a residual $\mathbf{r}_{k+1}$ and a matrix $\mathbf{c}_{k+1}$ in $\mathcal{G}_{k+1}$.

(a) Adopt (a) of Exercise 11.4 do this situation where $s > 1$.

Let $\mathbf{v}$ be an $n \times s$ matrix such that the columns of $\mathbf{v}$ span the Krylov subspace $\mathcal{K}_s(\mathbf{A}\Pi_1, \mathbf{A}\mathbf{r}'_k)$.

(b) Adopt (c) of Exercise 11.4: the $n \times s$ matrix $\mathbf{c}'_k$ spans $\mathcal{K}_s(\Pi_1\mathbf{A}, \Pi_1\mathbf{A}\mathbf{r}'_k)$

(c) We develop an algorithm for $\mathbf{v}$ the power basis of $\mathcal{K}_s(\mathbf{A}\Pi_1, \mathbf{A}\mathbf{r}'_k)$, that is, $\mathbf{v} = [\mathbf{v}_1, \ldots, \mathbf{v}_s]$ and the column vectors $\mathbf{v}_j = (\mathbf{A}\Pi_1)^{j-1}\mathbf{v}_1$, with $\mathbf{v}_1 \equiv \mathbf{A}\mathbf{r}'_k$, form the power bases of $\mathcal{K}_s(\mathbf{A}\Pi_1, \mathbf{A}\mathbf{r}'_k)$.

We consider the kth step of the algorithm. For ease of notation we drop the step index k. Note that this is in line with the implementation of the algorithm, where we allow old quantities to be replaced by the corresponding new ones. Put $\mathbf{c}' = [\mathbf{c}'_1, \ldots, \mathbf{c}'_s]$ and $\mathbf{u} = [\mathbf{u}'_1, \ldots, \mathbf{u}'_s]$. Note that the computation of the $\mathbf{v}_j$ as

$$\mathbf{r}'_k \rightsquigarrow \mathbf{v}_1 = \mathbf{A}\mathbf{r}'_k \rightsquigarrow \mathbf{c}'_1 = \Pi_1\mathbf{v}_1 \rightsquigarrow \mathbf{v}_2 \equiv \mathbf{A}\mathbf{c}'_1 \rightsquigarrow \ldots \rightsquigarrow \mathbf{c}'_s = \Pi_1\mathbf{v}_s \rightsquigarrow \mathbf{v}_{s+1} = \mathbf{A}\mathbf{c}'_s$$

requires $s+1$ MVs (multiplications of an n-vector by $\mathbf{A}$) and gives both $\mathbf{c}'$ and $\mathbf{A}\mathbf{c}'$ as a side product. Note that we compute $\mathbf{v}_{s+1}$ as well, since the computation of the next $\mathbf{c}$ requires $\mathbf{A}\mathbf{c}'_s$. Show that the $\mathbf{u}'_j$ can be computed as

$$\mathbf{r}'_k \rightsquigarrow \mathbf{u}'_1 = \Pi_0\mathbf{r}'_k \rightsquigarrow \mathbf{u}'_2 \equiv \Pi_0\mathbf{c}'_1 \rightsquigarrow \ldots \rightsquigarrow \mathbf{u}'_s = \Pi_0\mathbf{c}'_{s-1}.$$

Compare $\mathbf{c}'_{j+1} = \Pi_1\mathbf{v}_{j+1} = \Pi_1\mathbf{A}\mathbf{c}'_j$ and $\mathbf{u}'_{j+1} = \Pi_0\mathbf{c}'_j$. Show that these computations do not require new inner product, only new vector updates are needed.

Give an algorithm that relies on this power method approach.

(d) Show that if the s columns of the initial $\mathbf{u}_0$ ($k = 0$ here) form a basis of $\mathcal{K}_s(\mathbf{A}, \mathbf{r}_0)$ (whence the columns of $\mathbf{c}_0$ span $\mathcal{K}_s(\mathbf{A}, \mathbf{A}\mathbf{r}_0)$), then $\mathbf{r}_k$ belongs to $\mathcal{K}_{k(s+1)+1}(\mathbf{A}, \mathbf{r}_0)$. In particular, with $\mathbf{r}_k = p_k(\mathbf{A})\mathbf{v}$, we have that $\mathbf{v} \in \mathcal{K}_{ks+1}(\mathbf{A}, \mathbf{r}_0) \cap (\mathcal{K}_k(\mathbf{A}^*, \mathbf{R})^\perp$. Generally, $(\mathcal{K}_k(\mathbf{A}^*, \mathbf{R})^\perp$ has dimension $n - ks$, whence, generally, $\mathcal{K}_{ks+1}(\mathbf{A}, \mathbf{r}_0) \cap (\mathcal{K}_k(\mathbf{A}^*, \mathbf{R})^\perp$ has dimension one.

Investigate numerically the effect of starting with a Krylov basis for $\mathbf{u}_0$ versus a random $n \times s$ matrix $\mathbf{u}_0$.

(e) Give also an algorithm for $\mathbf{v}$ an orthonormal basis (an Arnoldi basis of $\mathcal{K}_s(\mathbf{A}\Pi_1, \mathbf{A}\mathbf{r}'_k)$). It is also possible to orthonormalise $\mathbf{c}'_k$ (an Arnoldi basis of $\mathcal{K}_s(\Pi_1\mathbf{A}, \Pi_1\mathbf{A}\mathbf{r}'_k)$). Make sure that you need only $s+1$ MVs to form $\mathbf{r}_{k+1}$, $\mathbf{x}_{k+1}$, $\mathbf{u}_{k+1}$, $\mathbf{c}_{k+1}$ from $\mathbf{r}_k$, $\mathbf{x}_k$, $\mathbf{u}_k$, $\mathbf{c}_k$. Keep the number of inner products limited.

(f) Compare these variants numerically. What is the effect of increasing s? What is the effect of a random $\mathbf{R}$ (or $\mathbf{R} = \mathbf{Q}$ with $\mathbf{Q}$ the $n \times s$ orthonormal component of the economical QR-decomposition of an $n \times s$ random matrix) versus a more structured choice (a Krylov basis for $\mathcal{K}_s(\mathbf{A}^*, \widetilde{\mathbf{r}}_0)$)?

B Saddle point problems

Let $\mathbf{A}$ be given in block form as

$$\mathbf{A} = \begin{bmatrix} \mathbf{F} & \mathbf{B} \\ \mathbf{B}^* & -\mathbf{C} \end{bmatrix}, \tag{11.8}$$

with both $\mathbf{F}$ and $\mathbf{C}$ Hermitian, $\mathbf{F}$ $m \times m$ positive definite and $\mathbf{C}$ $k \times k$ positive semi-definite. The matrix $\mathbf{B}$ is not square, $k \leq m$.

Applications

Matrices of this type show up in many applications. As we learnt in Exercise 9.1, some formulations of the least square problems lead to these matrices. We now discuss two other applications.

Exercise 11.6. **Quadratic Optimisation.** Let f and $g_1, \ldots, g_k$ be real-valued twice continuously differentiable functions defined on $\mathbb{C}^n$. Here, $k \leq n$. We are interested in an n-vector, say $\mathbf{x}^\star$, that minimises $f(\mathbf{x})$ for all n-vectors $\mathbf{x}$ that satisfy $g_j(\mathbf{x}) = 0$ for all $j = 1, \ldots, k$. $\mathbf{x}^\star$ solves the **minimisation problem**

$$\min f(\mathbf{x}) \quad \text{such that} \quad g_j(\mathbf{x}) = 0 \quad (j = 1, \ldots, k). \tag{11.9}$$

$\mu^\star \equiv f(\mathbf{x}^\star)$ is the minimising value.

(a) Put $\mathcal{G} \equiv \bigcap_j \{\mathbf{x} \in \mathbb{C}^n \,\big|\, g_j(\mathbf{x}) = 0\}$. Argue that

$$\mathcal{G} \;\text{ is tangent to the level curve }\; \{\mathbf{x} \in \mathbb{C}^n \,\big|\, f(\mathbf{x}) = \mu^\star\} \;\text{ at }\; \mathbf{x}^\star.$$

(b) In particular, (11.9) implies that the gradient $\nabla f(\mathbf{x}^\star)$ of f at $\mathbf{x}^\star$ is in the space that is orthogonal to $\mathcal{G}$ at $\mathbf{x}^\star$. Show that this space is spanned by the vectors $\nabla g_1(\mathbf{x}^\star), \ldots, \nabla g_k(\mathbf{x}^\star)$. Conclude that

$$\nabla f(\mathbf{x}) + \sum_{j=1}^{k} \lambda_j \nabla g_j(\mathbf{x}) = 0 \quad \text{and} \quad g_j(\mathbf{x}) = 0 \quad (j = 1, \ldots, k) \tag{11.10}$$

for $\mathbf{x} = \mathbf{x}^\star$ and certain k-vector $\lambda^\star = \lambda = (\lambda_1, \ldots, \lambda_k)^{\mathrm{T}}$. The λ_j are **Lagrange multipliers**. The equations in (11.10) are the **KKT conditions** (Karush-Kuhn-Tucker conditions) for the optimisation problem (11.9). Note that these equations characterise the solution $\mathbf{x}^\star$ of (11.9) in case they have a unique solution.

(c) Let $\mathbf{H}$ be an $n \times n$ Hermitian positive definite matrix, $\mathbf{q}$ an n-vector, α a real number, $\mathbf{A}$ an $n \times k$ matrix and b a k-vector. Consider the **quadratic optimisation** problem:

$$\min\left(\tfrac{1}{2}\mathbf{x}^*\mathbf{H}\mathbf{x} + \mathbf{q}^*\mathbf{x} + \alpha\right) \quad \text{with } \mathbf{x} \text{ such that} \quad \mathbf{A}^*\mathbf{x} = b. \tag{11.11}$$

With $f(\mathbf{x}) \equiv \frac{1}{2}\mathbf{x}^*\mathbf{H}\mathbf{x} + \mathbf{q}^*\mathbf{x} + \alpha$ and $g_j(\mathbf{x}) \equiv \mathbf{e}_j^*(\mathbf{A}^*\mathbf{x} - b)$, this problem is of the form (11.9). Show that the KKT conditions now take the form

$$\begin{bmatrix} \mathbf{H} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \lambda \end{bmatrix} = \begin{bmatrix} -\mathbf{q} \\ b \end{bmatrix}.$$

The situation as described above is a simplification of the optimisation problems that one encounters in practice. Functions can be defined on a subset (domain) of $\mathbb{C}^n$ only. Moreover, the restrictions usually include inequalities $h_i(\mathbf{x}) \leq 0$ $(i = 1, \ldots, m)$ as well as equalities $g_j(\mathbf{x}) = 0$ $(j = 1, \ldots, k)$ (with $k + m \leq n$). The Lagrange multipliers $\mu = (\mu_1, \ldots, \mu_m)^{\mathrm{T}}$ in

$$\nabla f(\mathbf{x}) + \sum_{j=1}^{k} \lambda_j \nabla g_j(\mathbf{x}) + \sum_{i=1}^{m} \mu_i \nabla h_i(\mathbf{x}) = 0$$

for the inequality restrictions are required to satisfy an inequality restriction as well: $\mu_i \geq 0$ all i.

The **Stokes equation system** (here in 2-d formulation)

$$\begin{cases} -\mu\left(\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}\right) + \dfrac{\partial p}{\partial x} + f = 0 \\ -\mu\left(\dfrac{\partial^2 v}{\partial x^2} + \dfrac{\partial^2 v}{\partial y^2}\right) + \dfrac{\partial p}{\partial y} + g = 0 \end{cases} \qquad \text{and} \quad \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$$

describes slow viscous flows. It is a simplified version of the Navier–Stokes equation system for modelling the flow of an incompressible (Newtonian) fluid (as air and water). The first equation represents the conservation of momentum, the second equation enforces conservation of mass (the incompressibility constrain): (u, v) is the wind field, p is the pressure, f and g are external forces as gravitation. The momentum equation in full Navier–Stokes contains the non-linear 'convection' term $((\nabla\vec{u}^*)^*\vec{u})$ from the motion of the underlying fluid. The term is quadratic (in the velocity) and can be neglected for 'slow' flows. With the Laplacian $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$, the gradient $\nabla = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y})^{\mathrm{T}}$ and divergence $\nabla^* = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y})$, the Stokes equation can be represented as

$$\begin{bmatrix} -\mu\Delta & 0 & \frac{\partial}{\partial x} \\ 0 & -\mu\Delta & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ p \end{bmatrix} = \begin{bmatrix} -f \\ -g \\ 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} -\mu\vec{\Delta} & \nabla \\ \nabla^* & 0 \end{bmatrix} \begin{bmatrix} \vec{u} \\ p \end{bmatrix} = \begin{bmatrix} -\vec{f} \\ 0 \end{bmatrix}.$$

$\vec{\Delta}$ is the 'vector Laplacian'. Discretization leads to a matrix as in (11.8) with $\mathbf{F}$ the discretised vector Laplacian and $\mathbf{B}$ the discretised gradient operator.

The Stokes equation can be viewed as the KKT conditions of a (energy minimising) minimisation problem

$$\iint \tfrac{1}{2}\mu(\|\nabla u\|_2^2 + \|\nabla v\|_2^2) + (f\frac{\partial u}{\partial x} + g\frac{\partial v}{\partial y})\ \mathrm{d}x\ \mathrm{d}y \quad \text{such that} \quad \nabla^*\vec{u} = 0.$$

The pressure p represent the Lagrange multiplier.

Exercise 11.7.

(a) Show that the least norm problem $\mathbf{Ax} = \mathbf{b}$ ($\mathbf{A}$ is $k \times n$, $k \leq n$, $\mathbf{A}$ full rank) as in Exercise 9.1(b) can be viewed as the KKT conditions of a quadratic minimisation problem.

(b) Show that the least square problem $\mathbf{Ax} = \mathbf{b}$ ($\mathbf{A}$ is $n \times k$, $k \leq n$, $\mathbf{A}$ full rank) as in Exercise 9.1(a) can be viewed as the KKT conditions of a quadratic minimisation problem. What are the Lagrange multipliers?

Conditioning

With $\mathbf{M}_S \equiv \mathbf{C} + \mathbf{B}^*\mathbf{F}^{-1}\mathbf{B}$, $-\mathbf{M}_S$ is the **Schur complement** and we have the **block LU-decomposition**

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{B}^*\mathbf{F}^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{F} & \mathbf{0} \\ \mathbf{0}^* & -\mathbf{M}_s \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{F}^{-1}\mathbf{B} \\ \mathbf{0}^* & \mathbf{I} \end{bmatrix}.$$

The Schur complement is useful to establish the stability (conditioning) of the matrix (see Exercise 11.8) and leads to effective preconditioners.

A matrix $\mathbf{A}$ of (11.8) is Hermitian, but indefinite (see (a) of Exercise 11.8). Note that the Schur complement $\mathbf{M}_S$ is negative semi-definite.

Exercise 11.8.

(a) Compute the eigenvalues of $\mathbf{A}$ in case $k = m = 1$: $\mathbf{F} = \alpha > 0$, $\mathbf{b} = \beta$, $\mathbf{C} = \gamma \geq 0$.

(b) Assume $\mathbf{C} = \mathbf{0}$. Note that $\mathbf{A}$ is singular if $\mathbf{B}^*$ does not have full column rank. The matrix $\mathbf{A}$ is non-singular if and only if the Schur complement is non-singular.

For a $\beta > 0$, prove that the following two statements are equivalent

$$\|\mathbf{B}^*\mathbf{F}^{-1}\mathbf{Bp}\|_2 \geq \beta\|\mathbf{p}\|_2 \qquad (\mathbf{p} \in \mathbb{C}^k)$$

and (the inf-sup condition or Babushka–Brezzi condition) for every $\mathbf{p} \in \mathbb{C}^k$ there is a $\mathbf{v} \in \mathbb{C}^m$ such that

$$\mathbf{v}^*\mathbf{Bp} \geq \beta\,\|\mathbf{p}\|_2\,\sqrt{\mathbf{v}^*\mathbf{Fv}}.$$

The inf-sup condition plays an important role in mixed finite element methods for the Stokes equation.

Show that $\|\mathbf{A}^{-1}\|_2 \leq (1 + \|\mathbf{F}^{-1}\mathbf{B}\|_2)\max(\|\mathbf{F}^{-1}\|_2, 1/\beta)$.

(c) Prove that the matrix $\mathbf{A}$ is non-singular if $\mathbf{C}$ is positive definite. $\mathbf{C}$ is often introduced to "stabilise" the matrix, i.e., the make an approximate matrix that is 'well conditioned' (as $\mathbf{C} = \tau^2\mathbf{I}$ in demped least square problems).

Preconditioning

As a preconditioner for the saddle point problem, consider the matrix

$$\mathbf{P} = \begin{bmatrix} \mathbf{F} & \mathbf{B} \\ \mathbf{O}^* & -\mathbf{M}_S \end{bmatrix} \tag{11.12}$$

This leads to the right-preconditioned matrix

$$\mathbf{A}\mathbf{P}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{B}^*\mathbf{F}^{-1} & \mathbf{I} \end{bmatrix}.$$

Exercise 11.9. Let $\mathbf{v}$ be any nonzero vector.

(a) Compute $\mathbf{A}\mathbf{P}^{-1}\mathbf{v}$ and $(\mathbf{A}\mathbf{P}^{-1})^2\mathbf{v}$. Show that $\mathbf{v}$ can be expressed as a linear combination of $\mathbf{A}\mathbf{P}^{-1}\mathbf{v}$ and $(\mathbf{A}\mathbf{P}^{-1})^2\mathbf{v}$.

(b) Explain why this implies that GMRES applied to a right-preconditioned system saddle point system with (11.12) as preconditioner must find the exact solution in at most two iterations.

The preconditioner in (11.12) is not Hermitian and leads to a skew preconditioned system. This may make the method (preconditioned GMRES) sensitive to perturbations. In practise, systems involving $\mathbf{F}$ can not be exactly solved (for instance, if $\mathbf{F}$ is a discrete Laplacian). An Hermitian preconditioner may not be helpful either: the preconditioned system may also have an ill conditioned basis of eigenvectors. See also the discussion following Exercise 10.3 and Exercise 10.4. We are interested in positive definite preconditioners. Exercise 10.4 describes a way to incorporate such a preconditioner in MINRES. A positive definite preconditioner for saddle point systems is the block diagonal matrix

$$\mathbf{P} = \begin{bmatrix} \mathbf{F} & \mathbf{O} \\ \mathbf{O}^* & \mathbf{M}_S \end{bmatrix}. \tag{11.13}$$

Exercise 11.10.

(a) Assume that $\mathbf{C} = \mathbf{0}$. Show that in this case the preconditioned matrix $\mathbf{A}\mathbf{P}^{-1}$ has three distinct eigenvalues: 1, $\frac{1}{2} + \frac{1}{2}\sqrt{5}$, and $\frac{1}{2} - \frac{1}{2}\sqrt{5}$.
Hint: solve the generalised eigenvalue problem

$$\begin{bmatrix} \mathbf{F} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{x}_t \\ \mathbf{x}_b \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{F} & \mathbf{O} \\ \mathbf{O}^* & \mathbf{M}_S \end{bmatrix} \begin{bmatrix} \mathbf{x}_t \\ \mathbf{x}_b \end{bmatrix}$$

by first eliminating $\mathbf{x}_t$.

(b) Assume again that $\mathbf{C} = \mathbf{0}$. Explain why MINRES applied to a saddle-point system, preconditioned with (11.13) must find the exact solution in at most three iterations.