Chapter 6

Supplementary Note

6.1 Iterative method

Given a symmetric positive definite $n \times n$ matrix A, we consider minimization problem: Given a quadratic functional

$$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x} + c$$

find the minimizer of ϕ.

Theorem 6.1.1. $\mathbf{x}_0$ *is a minimizer of* ϕ *if and only if* $A\mathbf{x}_0 = \mathbf{b}$.

6.1.1 Method of steepest descent

How to find the minimizer ? We start with an arbitrary initial guess $\mathbf{x}^0$. We try to find the next approximation in the form

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \tau_k \mathbf{d}^k \tag{6.1}$$

d^k is called search direction where τ_k is chosen to minimize, or reduce $\phi(\mathbf{x})$ on some interval near $\mathbf{x}^k$ in that direction. We need to choose $\mathbf{d}^k$ and τ_k. We know

Theorem 6.1.2. $\mathbf{d}^k = -\nabla\phi(\mathbf{x}^k) = \mathbf{b} - A\mathbf{x}^k$ *is the direction of steepest descent.*

To determine the parameter τ_k, we see

$$\phi(\mathbf{x}^k + \tau_k\mathbf{d}^k) = \frac{1}{2}\tau_k^2\mathbf{d}^{kT}A\mathbf{d}^k + \tau\mathbf{d}^T\nabla\phi(\mathbf{x}^k) + \hat{c} = \frac{1}{2}\tau^2(A\mathbf{d}^k, \mathbf{d}^k) - \tau(\mathbf{d}^k, \mathbf{d}^k) + \hat{c}.$$

Thus $\min_\tau \phi(\mathbf{x}^k + \tau_k\mathbf{d}^k)$ is obtained when

$$\frac{d}{d\tau}\phi(\mathbf{x}^k + \tau_k\mathbf{d}^k) = \tau(A\mathbf{d}^k, \mathbf{d}^k) - (\mathbf{d}^k, \mathbf{d}^k) = 0.$$

Thus, $\tau_k = -\frac{(\mathbf{d}^k,\mathbf{d}^k)}{(A\mathbf{d}^k,\mathbf{d}^k)}$. Also, the next search direction is given by

$$\mathbf{d}^{k+1} = \mathbf{b} - A\mathbf{x}^{k+1} = \mathbf{b} - A(\mathbf{x}^k + \tau_k\mathbf{d}^k) = \mathbf{d}^k - \tau_k A\mathbf{d}^k.$$

Now the method of steepest descent is described as follows:

$$\begin{aligned}
\mathbf{d}^k &= \mathbf{b} - A\mathbf{x}^k \\
\mathbf{x}^{k+1} &= \mathbf{x}^k + \tau_k\mathbf{d}^k \\
\tau_k &= \frac{(\mathbf{d}^k,\mathbf{d}^k)}{(A\mathbf{d}^k,\mathbf{d}^k)} \\
\mathbf{d}^{k+1} &= \mathbf{d}^k - \tau_k A\mathbf{d}^k.
\end{aligned}$$

6.1.2 Convergence analysis

We introduce a norm $(\cdot,\cdot)_A$ by $(\mathbf{x},\mathbf{y})_A = (A\mathbf{x},\mathbf{y})$. When A is symmetric, positive definite, $(\cdot,\cdot)_A$ becomes a true norm (called energy norm) on $\mathbb{R}n$.

Theorem 6.1.3. *We have*

$$\|\mathbf{x}^k - \mathbf{x}\|_A \le \left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^k \|\mathbf{x}^0 - \mathbf{x}\|_A,$$

where $\kappa(A)$ is the spectral condition number of A. Furthermore the number of iteration to reduce the error by a factor ϵ is

$$N \le \frac{1}{2}\kappa(A)\ln(1/\epsilon) + 1.$$

6.1.3 Conjugate gradient

The steepest descent is very slow. Hence we need another direction. The idea is to choose a new direction so that it is A-orthogonal to previous direction. Let $\mathbf{x}^0 = 0$, $\mathbf{d}^0 = \mathbf{r}^0 = \mathbf{b}$ and

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k\mathbf{d}^k \tag{6.2}$$

$$\mathbf{r}^k = \mathbf{b} - A\mathbf{x}^k. \tag{6.3}$$

We choose, as in the method of steepest descent, α_k so that $\phi(\mathbf{x}^k + \alpha_k\mathbf{d}^k)$ is minimized. Thus,

$$\alpha_k = (\mathbf{d}^k,\mathbf{r}^k)/(A\mathbf{d}^k,\mathbf{d}^k) \tag{6.4}$$

It also makes the residual $\mathbf{r}^{k+1}$ to be orthogonal to the search direction $\mathbf{d}^k$,

$$(\mathbf{d}^k,\mathbf{r}^{k+1}) = (\mathbf{d}^k,\mathbf{b} - A\mathbf{x}^{k+1}) = (\mathbf{d}^k,\mathbf{b} - A\mathbf{x}^k - \alpha_k A\mathbf{d}^k) = 0. \tag{6.5}$$

Now let us determine new direction in the form: $\mathbf{d}^{k+1} = \mathbf{r}^{k+1} - \beta_k \mathbf{d}^k$. (If $\beta_k = 0$, it is steepest descent.) Assuming

$$(A\mathbf{d}^j, \mathbf{d}^k) = 0, \quad j \le k-1, \tag{6.6}$$

we choose $\mathbf{d}^{k+1}$ so that it is orthogonal to $\mathbf{d}^k$:

$$0 = (A\mathbf{d}^{k+1}, \mathbf{d}^k) = (A\mathbf{r}^{k+1} - \beta_k A\mathbf{d}^k, \mathbf{d}^k). \tag{6.7}$$

Thus we obtain $\beta_k = (A\mathbf{d}^k, \mathbf{r}^{k+1})/(A\mathbf{d}^k, \mathbf{d}^k)$.

Now let

$$V_k = \mathrm{SPAN}\{\mathbf{b}, A\mathbf{b}, \cdots, A^{k-1}\mathbf{b}\}$$

then it is easy to see that

$$V_k = \mathrm{SPAN}\{\mathbf{d}^0, \mathbf{d}^1, \cdots, \mathbf{d}^{k-1}\}.$$

We claim

Lemma 6.1.4.

$$\mathbf{d}^{k+1} \perp V_{k+1} \textit{ with respect to } (A\cdot, \cdot) \tag{6.8}$$

$$\mathbf{r}^{k+1} \perp V_k \textit{ with respect to } (A\cdot, \cdot) \tag{6.9}$$

Proof. Since the first relation is obvious from the construction of $\mathbf{d}^k$, it suffices to show

$$(A\mathbf{d}^j, \mathbf{r}^{k+1}) = 0, \quad j \le k-1.$$

We see

$$(A\mathbf{d}^j, \mathbf{r}^{k+1}) = (A\mathbf{d}^j, \mathbf{d}^{k+1}) - \beta_k (A\mathbf{d}^j, \mathbf{d}^k) = 0$$

by induction (6.6). □

Let $\mathbf{e}^k = \mathbf{x}^k - \mathbf{x}$. Then from (6.5), we see that $(\mathbf{d}^k, A(\mathbf{x} - \mathbf{x}^{k+1})) = 0$ or $(A\mathbf{e}^k, \mathbf{d}^{k-1}) = 0$.

Thus,

$$\mathbf{e}^k = \mathbf{x}^k - \mathbf{x} \perp V_k \text{ with respect to } (A\cdot, \cdot). \tag{6.10}$$

The algorithm is

$$\mathbf{d}^0 = \mathbf{r}^0, \quad \mathbf{x}^0 = 0$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{d}^k, \quad \alpha_k = (\mathbf{d}^k, \mathbf{r}^k)/(A\mathbf{d}^k, \mathbf{d}^k) \tag{6.11}$$

$$\mathbf{r}^{k+1} = \mathbf{b} - A\mathbf{x}^{k+1} = \mathbf{r}^k - \alpha_k A\mathbf{x}^k \tag{6.12}$$

$$\mathbf{d}^{k+1} = \mathbf{r}^{k+1} - \beta_k \mathbf{d}^k \quad \beta_k = (A\mathbf{d}^k, \mathbf{r}^{k+1})/(A\mathbf{d}^k, \mathbf{d}^k). \tag{6.13}$$

Note that since $\mathbf{r}^{k+1} = \mathbf{r}^k - \alpha_k A\mathbf{d}^k$, only one evaluation of A is necessary and no need to estimate β_k.

Remark 6.1.5. One can check that

$$\alpha_k = (\mathbf{r}^k, \mathbf{r}^k)/(A\mathbf{d}^k, \mathbf{d}^k)$$

and

$$\beta_k = -(\mathbf{r}^{k+1}, \mathbf{r}^{k+1})/(\mathbf{r}^k, \mathbf{r}^k).$$

6.1.4 Error analysis

By (6.10) we have

$$(A\mathbf{e}^k, \mathbf{e}^k) = (A\mathbf{e}^k, \mathbf{x}^k - \mathbf{y} + \mathbf{y} - \mathbf{x}) = (A\mathbf{e}^k, \mathbf{y} - \mathbf{x}), \ \forall \mathbf{y} \in V_k \tag{6.14}$$
$$\tag{6.15}$$

and Cauchy Schwarz inequality,

$$(A\mathbf{e}^k, \mathbf{e}^k) \le (A(\mathbf{x} - \mathbf{y}), \mathbf{x} - \mathbf{y}), \ \forall \mathbf{y} \in V_k. \tag{6.16}$$

Since $y \in V_k$,

$$\mathbf{y} = P_k(A)\mathbf{b},$$

for some polynomial $P_{k-1}(t)$ of degree $k-1$. Hence

$$\begin{aligned}(A(\mathbf{x} - \mathbf{y}), \mathbf{x} - \mathbf{y}) &= (A(I - P_{k-1}(A))\mathbf{b}, (I - P_{k-1}(A))\mathbf{b}) \\ &\le \|I - P_{k-1}(A)A\|^2 (A\mathbf{x}, \mathbf{x}).\end{aligned}$$

Thus

$$(A\mathbf{e}^k, \mathbf{e}^k) \le \min_{Q_k \in P_k} \|Q_k(A)\|^2 (A\mathbf{x}, \mathbf{x}),$$

where $\|\cdot\|$ is the matrix norm and $Q_k(t)$ is any polynomial of degree k with $Q(0) = 1$. Let $\tilde{Q}_k(t)$ be such that $\tilde{Q}_k(1) = 1$ and set $Q_k(t) = \tilde{Q}_k(1 - \frac{2}{\lambda_N + \lambda_1} t)$. Then $Q_k(0) = 1$ and $Q_k(A) = \tilde{Q}_k(M)$ where $M = I - \frac{2}{\lambda_N + \lambda_1} A$. Then we see that $\sigma(M) \subset [-\rho, \rho]$, where

$$\rho = \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1} < 1.$$

Thus

$$\min_{Q_k(0)=1} \|Q_k(A)\| = \min_{\tilde{Q}_k(1)=1} \|\tilde{Q}_k(M)\| \tag{6.17}$$
$$= \min_{\tilde{Q}_k(1)=1,} \max_{\lambda \in \sigma(M)} |\tilde{Q}_k(\lambda)| \tag{6.18}$$
$$= \min_{\tilde{Q}_k(1)=1} \max_{\lambda \in [-\rho, \rho]} |\tilde{Q}_k(\lambda)| \tag{6.19}$$

The best choice is given by Chebyshev polynomial on $[-\rho, \rho]$ which is $\tilde{Q}_k(x) = C_k(\frac{x}{\rho})/C_k(\frac{1}{\rho})$. The minimum value can be seen to be

$$2\left(\frac{\sqrt{K} - 1}{\sqrt{K} + 1}\right)^n$$

where $K = \frac{\lambda_N}{\lambda_1}$.

Remark 6.1.6. Conjugate gradient method ends in finite steps: If we choose P_N so that $1-\lambda_j P_{N-1}(\lambda_j)=0, \quad \forall \lambda_j \in \sigma(A)$, then $A(\mathbf{e}^N,\mathbf{e}^N)=0$ and hence $\mathbf{x}^N=\mathbf{x}$.

Remark 6.1.7. If the eigenvalues are accumulated near λ_1, then let $Q(t)=Q^1(t)Q^2(t)$, where $Q^2(t)$ is a polynomial of lower degree which is bounded by small constant C.

$$|Q| \le \max_{\sigma(M)} |Q^1(t)| \max_{\sigma(M)} |Q^2| \le \max_{[-\tau,\tau]} |Q^1(t)| \le C \left(\frac{\lambda_1-\tilde{\lambda}_0}{\lambda_1+\tilde{\lambda}_0}\right)^{n-k} \le \left(\frac{\lambda_1-\lambda_0}{\lambda_1+\lambda_0}\right)^{n}$$

for large n.

6.1.5 Preconditioning

Consider

$$R^{-1}A\mathbf{x}=R^{-1}\mathbf{b}=\tilde{\mathbf{b}}.$$

we introduce an inner product $[\cdot,\cdot]$ as either $(A\cdot,\cdot)$ or $(R\cdot,\cdot)$. Then the operator $R^{-1}A$ is symmetric with respect to $[\cdot,\cdot]$,i.e.

$$[R^{-1}A\mathbf{x},\mathbf{y}]=[\mathbf{x},R^{-1}A\mathbf{y}].$$

R^{-1} is called a preconditioner for A. Two properties of preconditioner is desirable:

(1) The action of R^{-1} on an arbitrary vector is in some sense "easy" to compute.

(2) Since A and R are both SPD, there exist $\tilde{\lambda}_0, \tilde{\lambda}_N$ such that

$$\tilde{\lambda}_0(R\mathbf{x},\mathbf{x}) \le (A\mathbf{x},\mathbf{x}) \le \tilde{\lambda}_N(R\mathbf{x},\mathbf{x}).$$

The condition number of $R^{-1}A=\tilde{\lambda}_N/\tilde{\lambda}_0$ should be smaller than that of A.

Application to Conjugate Gradient Method

One could directly apply cg-method to the preconditioned system. But it is sometimes hard to estimate the condition number of the first type of preconditioner. Thus, we consider an alternative way:

The idea is to apply the cg with respect to new inner product: With $\tilde{\mathbf{r}}^0=\tilde{\mathbf{d}}^0=\tilde{\mathbf{b}}-R^{-1}A\mathbf{x}^0$, the algorithm is

$$\begin{aligned}
\mathbf{x}^{k+1} &= \mathbf{x}^k+\alpha_k\tilde{\mathbf{d}}^k, \alpha_k=[\tilde{\mathbf{d}}^k,\tilde{\mathbf{r}}^k]/[R^{-1}A\tilde{\mathbf{d}}^k,\tilde{\mathbf{d}}^k]\\
\tilde{\mathbf{r}}^{k+1} &= \tilde{\mathbf{r}}^k-\alpha_k R^{-1}A\tilde{\mathbf{d}}^k\\
\tilde{\mathbf{d}}^{k+1} &= \tilde{\mathbf{r}}^{k+1}-\beta_k\tilde{\mathbf{d}}^k, \beta_k=[R^{-1}A\tilde{\mathbf{d}}^k,\tilde{\mathbf{r}}^{k+1}]/[R^{-1}A\tilde{\mathbf{d}}^k,\tilde{\mathbf{d}}^k]
\end{aligned}$$

With $[\cdot,\cdot] = (R\cdot,\cdot)$, the algorithm becomes

$$\begin{aligned}
\mathbf{x}^{k+1} &= \mathbf{x}^k + \alpha_k \tilde{\mathbf{d}}^k, \alpha_k = (R\tilde{\mathbf{d}}^k, \tilde{\mathbf{r}}^k)/(A\tilde{\mathbf{d}}^k, \tilde{\mathbf{d}}^k) \\
\tilde{\mathbf{r}}^{k+1} &= \tilde{\mathbf{r}}^k - \alpha_k R^{-1} A\tilde{\mathbf{d}}^k \\
\tilde{\mathbf{d}}^{k+1} &= \tilde{\mathbf{r}}^{k+1} - \beta_k \tilde{\mathbf{d}}^k, \beta_k = (A\tilde{\mathbf{d}}^k, \tilde{\mathbf{r}}^{k+1})/(A\tilde{\mathbf{d}}^k, \tilde{\mathbf{d}}^k)
\end{aligned}$$

This algorithm could be disastrous because we need to evaluate $R\mathbf{d}^k$ at each step. To avoid this, we write the algorithm in an equivalent form

$$\begin{aligned}
\mathbf{x}^{k+1} &= \mathbf{x}^k + \alpha_k \tilde{\mathbf{d}}^k, \alpha_k = (\tilde{\mathbf{d}}^k, R\mathbf{z}^k)/(A\tilde{\mathbf{d}}^k, \tilde{\mathbf{d}}^k) \\
\mathbf{r}^{k+1} &= \mathbf{r}^k - \alpha_k A\tilde{\mathbf{d}}^k \\
\mathbf{z}^{k+1} &= R^{-1}\mathbf{r}^{k+1} \\
\tilde{\mathbf{d}}^{k+1} &= \mathbf{z}^{k+1} - \beta_k \tilde{\mathbf{d}}^k, \beta_k = (A\tilde{\mathbf{d}}^k, \mathbf{z}^{k+1})/(A\tilde{\mathbf{d}}^k, \tilde{\mathbf{d}}^k)
\end{aligned}$$

and change the starting value:

With $\mathbf{r}^0 = \mathbf{b} - A\mathbf{x}^0$, $\mathbf{d}^0 = \mathbf{z}^0 = R^{-1}\mathbf{r}^0$,

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{d}^k, \alpha_k = (\mathbf{d}^k, \mathbf{r}^k)/(A\mathbf{d}^k, \mathbf{d}^k) \tag{6.20}$$
$$\mathbf{r}^{k+1} = \mathbf{r}^k - \alpha_k A\mathbf{d}^k \tag{6.21}$$
$$\mathbf{z}^{k+1} = R^{-1}\mathbf{r}^{k+1} \tag{6.22}$$
$$\mathbf{d}^{k+1} = \mathbf{z}^{k+1} - \beta_k \mathbf{d}^k, \beta_k = (A\mathbf{d}^k, \mathbf{z}^{k+1})/(A\mathbf{d}^k, \mathbf{d}^k) \tag{6.23}$$

This is the final algorithm where only one evaluation of A and R^{-1} is involved in each iteration.

Preconditioned iterative method

Consider an iterative method to solve $R^{-1}A\mathbf{x} = R^{-1}\mathbf{b} = \tilde{\mathbf{b}}$. We change it to the form

$$\mathbf{x} = \mathbf{x} - R^{-1}A\mathbf{x} + \tilde{\mathbf{b}}.$$

Hence we obtain an iterative method of the form

$$\mathbf{x}^{k+1} = M\mathbf{x}^k + \tilde{\mathbf{b}}, \tag{6.24}$$

where $M = I - R^{-1}A$. This will be convergent if $\rho(M) = \rho < 1$.

Lemma 6.1.8. *The condition number is* $\kappa(R^{-1}A) = \frac{1+\rho}{1-\rho}$ *iff* $\rho(M) = \rho < 1$.

Proof. Since M is symmetric ,

$$-\rho(R\mathbf{y}, \mathbf{y}) \le (RM\mathbf{y}, \mathbf{y}) \le \rho(R\mathbf{y}, \mathbf{y}).$$

This is equivalent to

$$\begin{aligned}
-\rho(R\mathbf{y}, \mathbf{y}) &\le (A\mathbf{y}, \mathbf{y}) - (R\mathbf{y}, \mathbf{y}) \le \rho(R\mathbf{y}, \mathbf{y}) \\
(1-\rho)(R\mathbf{y}, \mathbf{y}) &\le (A\mathbf{y}, \mathbf{y}) \le (1+\rho)(R\mathbf{y}, \mathbf{y}).
\end{aligned}$$

□

6.2 Linear Algebra

Theorem 6.2.1 (Schur)**.** *If $M \in \mathbb{C}^{n,n}$, then $\exists$ a unitary matrix U such that $U^*MU = T$, where T is upper triangular.*

Proof. Let λ_1 be an eigenvalue of M and $\mathbf{u}_1$ be a corresponding eigenvector(there exists at lest one) with $u_1 \geq 0$ and $\mathbf{u}_1^*\mathbf{u}_1 = \|\mathbf{u}_1\|^2 = 1$ so that $M\mathbf{u}_1 = \lambda\mathbf{u}_1$. Let $\mathbf{y}_2, \cdots, \mathbf{y}_n$ be a set of orthonormal vector which in turn are orthogonal to $\mathbf{u}_1$. Let $U_1 = (\mathbf{u}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n)$. Then U_1 is unitary and

$$MU_1 = U_1T_1$$

where the first column of T is $(\lambda_1, 0, \cdots, 0)$. Then we have

$$U_1^*MU_1 = T_1 = \begin{pmatrix} \lambda_1 & * \\ 0 & M_1 \end{pmatrix}.$$

Repeat the same process to M_2 to obtain $(n-1)\times(n-1)$ unitary matrix U_2 such that

$$U_2^*M_1U_2 = T_2 = \begin{pmatrix} \lambda_2 & * \\ 0 & M_2 \end{pmatrix}.$$

Let

$$V_2 = \begin{pmatrix} 1 & 0 \\ 0 & U_2 \end{pmatrix}$$

Then

$$V_2^*U_1^*M_1U_1V_2 = T_2' = \begin{pmatrix} \lambda_1 & * & * \\ 0 & \lambda_2 & * \\ 0 & M_3 & \end{pmatrix}.$$

Repeat the same process until we obtain the desired decomposition. The eigenvalues of U are clearly those of T obtained in this process. ☐

Theorem 6.2.2 (Singular value decomposition)**.** *If $A \in \mathbb{R}^{m\times n}$ then there exists orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ such that*

$$U^tAV = \Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_p) \tag{6.25}$$

where Σ is $m \times n$ matrix and $p = \min(m,n)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$.

Proof. Note that $B = AA^t \in \mathbb{R}^{m\times m}$ is real, symmetric matrix which is nonnegative definite. Thus B has orthonormal eigenvectors $\mathbf{u}_1, \cdots, \mathbf{u}_m$ with nonnegative eigenvalues λ_i. Define $\sigma_i = \sqrt{\lambda_i}$ and $U = (\mathbf{u}_1, \cdots, \mathbf{u}_m)$. Define $F = A^tU \in \mathbb{R}^{n\times m}$ and let

$$F = \left(\mathbf{f}_1, \cdots, \mathbf{f}_m\right), \quad \text{then } F^t = \begin{pmatrix} \mathbf{f}_1^t \\ \vdots \\ \mathbf{f}_m^t \end{pmatrix}$$

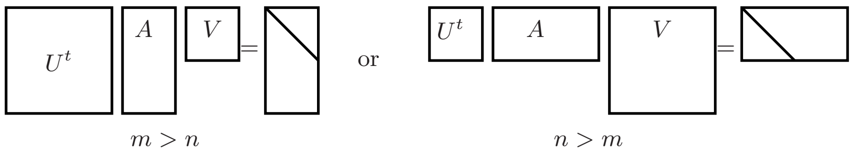

Figure 6.1: Σ of SVD

Observe that

$$F^t F = \mathrm{diag}(\sigma_i^2), \quad \mathbf{f}_i \cdot \mathbf{f}_j = \sigma_i^2 \delta_{ij}, i, j \le m.$$

The (k, k) entry of this equality asserts that k-th column of F(call it $\mathbf{f}_k$) has norm $\|\mathbf{f}_k\| = \sigma_k$. Furthermore, the off-diagonal elements of this equality asserts that distinct columns of F are orthogonal. Pick $r > 0$ such that $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0$ and $\sigma_{r+1} = 0$. Then $\{\mathbf{f}_i\}_{i=1}^r$ are the nonzero orthogonal vectors and

$$\mathbf{v}_i = \frac{1}{\sigma_i}\mathbf{f}_i, \quad i = 1, \cdots, r$$

defines an orthonormal set of vectors. We may expand $\{\mathbf{v}_i\}_{i=1}^r$ to an orthonormal basis of $\mathbb{R}^n$ by appending vectors $\{\mathbf{v}_{r+1}, \cdots, \mathbf{v}_n\}$. Now define the orthogonal matrix $V \in \mathbb{R}^{n \times n}$ by $V = (\mathbf{v}_1, \cdots, \mathbf{v}_n)$ and observe $F = V \cdot \mathrm{diag}(\sigma_i)$. Now

$$U^t A = F^t = \mathrm{diag}(\sigma_i) V^t \Rightarrow U^t A V = \mathrm{diag}_{i=1,\cdots,p}(\sigma_i).$$

Thus Σ is $m \times n$ (almost diagonal) matrix described above. □

Remark 6.2.3. (1) The Schur decomposition may be written

$$A = UTU^t, \quad A \in \mathbb{R}^{n \times n}$$

The Singular value decomposition may be written

$$A = U\Sigma V^t, \quad A \in \mathbb{R}^{m \times n}$$

Both of them are unitary transformations.

(2) When A is square the singular values and eigenvalues are not directly related in general. Let $= \begin{bmatrix} 1 & a \\ 0 & 5 \end{bmatrix}$ which is already in Schur form. The eigenvalues are $1, 5$. However, the singular values $\sigma_1 \to \infty$ and $\sigma_2 \to 0$ as $a \to \infty$.

(3) The 2-norm of any matrix $A \in \mathbb{R}^{m\times n}$ is given by the largest singular value: $\|A\|_2 = \sigma_1$. Verification: By orthogonality we see $\|U\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for any orthogonal matrix. Hence

$$\|A\|_2 = \max_{\mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \max \frac{\|U\Sigma V^t\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \max \frac{\|\Sigma V^t\mathbf{x}\|_2}{\|V^t\mathbf{x}\|_2} = \max_{\mathbf{y}\neq 0} \frac{\|\Sigma\mathbf{y}\|_2}{\|\mathbf{y}\|_2}.$$

6.3 projections

A matrix $P \in \mathbb{R}^{n\times n}$ is an orthogonal projection onto a subspace $S \subset \mathbb{R}^n$ if

(1) $Range(P) \subset S$

(2) $P^2 = P$

(3) $P^t = P$

In this case $I - P$ is also an orthogonal projection (onto $S^\perp$).

Example 6.3.1. (1) $P = \frac{1}{\|V\|^2} VV^t$ is an orthogonal projection onto $S = Span(V)$.

(2) Let the columns of $V = (\mathbf{v}_1, \cdots, \mathbf{v}_k)$ be orthonormal. Then $P = VV^t$ is an orthogonal projection onto $Span(\mathbf{v}_1, \cdots, \mathbf{v}_k)$. In fact, $P\mathbf{x} = \sum_{i=1}^k (\mathbf{v}_i^t\mathbf{x})\mathbf{v}_i$.

6.4 Pseudo inverse

If A is $m \times n$ matrix, what problems do we have in defining the "inverse" of A ?

- It may not be **one-to-one**
- It may not be **onto**. $(Ran(A) \neq \mathbb{R}^m)$

What if we restrict it to a subspace of $\mathbb{R}^n$? In fact one can find subspaces $S_1 \subset \mathbb{R}^n$ and $S_2 \subset \mathbb{R}^m$ so that A is one-to-one and onto $S_1 \to S_2$. To show how this can be done, we need some projections: Let P be the orthogonal projection onto $Ker(A)^\perp$. Then $I - P$ is the orthogonal projection onto $N(A) := Ker(A)$. Hence $\mathbf{x} - P\mathbf{x} \in Ker(A)$ so that $A\mathbf{x} = AP\mathbf{x}$. So we can imagine the action of A as

(1) a projection P and

(2) a transformation by A onto the range of A.

Now restrict this action only to $ker(A)^{\perp}$. Now $A : Ker(A)^{\perp} \to Rang(A)$ is one-to-one and onto: Suppose $A\mathbf{x}_1 = A\mathbf{x}_2$, for $\mathbf{x}_1, \mathbf{x}_2 \in Ker(A)^{\perp}$. Then $\mathbf{x}_1 - \mathbf{x}_2 \in Ker(A)^{\perp}$, Also, $\mathbf{x}_1 - \mathbf{x}_2 \in Ker(A)$. Hence $\mathbf{x}_1 - \mathbf{x}_2 = 0$. Now A can be thought of a the composition of a projection and an invertible transform.

Singular value decomposition provides a way of constructing such pseudo inverse(denoted by A^+)

$$A = U\Sigma V^t = (U_r : \bar{U}_r) \begin{pmatrix} \text{diag}(\sigma_i) & 0 \\ 0 & 0 \end{pmatrix} (V_r : \bar{V}_r)^t = U_r \Sigma_r V_r^t = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^t.$$

Indeed, V_r is the projection onto $Ker(A)^{\perp}$. Since the columns of V are orthogonal, we have

$$A\mathbf{v}_j = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i (\mathbf{v}_i^t \mathbf{v}_j) = 0, \quad j = r+1, \cdots, n.$$

Hence $\{\mathbf{v}_{r+1}, \cdots, \mathbf{v}_n\}$ forms a basis for $Ker(A)$ and $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ forms a basis for $Ker(A)^{\perp}$. Thus

$$P = V_r V_r^t$$

is a $n \times n$ orthogonal matrix which provides a projection onto $Ker(A)^{\perp}$. Similarly, $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ forms a basis for $Ran(A)$ so that $Q = U_r U_r^t$ is an orthogonal projection onto $Ran(A)$.

$$\begin{tikzcd} \mathbb{R}^n \arrow[r, "A"] \arrow[d, "P"'] & \mathbb{R}^m \arrow[d, "Q"] \\ N(A)^+ \arrow[r] & R(A) \end{tikzcd}$$

Figure 6.2: Pseudo Inverse

Transform back to $Ker(A)^{\perp}$

Since Q is the projection onto the range of A, the equation

$$A\mathbf{x} = Q\mathbf{y}$$

has a unique solution for any $\mathbf{y} \in \mathbb{R}^n$. In fact,

$$U_r \Sigma_r V_r^t \mathbf{x} = U_r U_r^t \mathbf{y}$$

implies

$$U_r(\Sigma_r V_r^t \mathbf{x} - U_r^t \mathbf{y}) = 0.$$

This is a linear combination of columns of U_r which are linearly independent. Hence we must have

$$\Sigma_r V_r^t \mathbf{x} = U_r^t \mathbf{y} \Rightarrow V_r^t \mathbf{x} = \Sigma_r^{-1} U_r^t \mathbf{y}$$

so that

$$(V_r V_r^t)\mathbf{x} = (V_r \Sigma_r^{-1} U_r^t)\mathbf{y}.$$

Since $V_r V_r^t$ is a projection ont $Ker(A)^\perp$ and $\mathbf{x} \in Ker(A)^\perp$, we obtain

$$\mathbf{x} = (V_r \Sigma_r^{-1} U_r^t)\mathbf{y}.$$

Hence we can define

$$A^+ = V_r \Sigma_r^{-1} U_r^t$$

or equivalently

$$A^+ = (V_r : \bar{V}_r) \begin{pmatrix} \mathrm{diag}(\frac{1}{\sigma_i}) & 0 \\ 0 & 0 \end{pmatrix} (U_r : \bar{U}_r^t) = V \Sigma^+ U^t.$$

Also, Σ^+ is the pseudo inverse of Σ.(Check it)

Remark 6.4.1. (1) $AA^+ \neq I$ in general. But $A^+A = V_r V_r^t = P$ which acts like I on $Ker(A)^\perp$. Likewise, $AA^+ = U_r U_r^t = Q$ acts like I on $Ran(A)$.

(2) A^+ provides the solution to the minimal least square problem:

Find a minimal($\|\mathbf{x}\|$ is minimal) solution $\mathbf{x} \in \mathbb{R}^n$ of such that

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\| \text{ with } \|\mathbf{x}\| \text{ is minimal.}$$

(3) The SVD is one way of constructing the pseudoinverse of A. Other ways are discussed in G. Peters and J. H. Wilkinson, "The least squares problem and pseudoinverses", Computer Jour, 13. pp 309-316(1970).

6.5 Least square problem

We consider the following problem: Find $\mathbf{x} \in \mathbb{R}^n$ such that

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2.$$

The solution always exists, but may not be unique. Every solution satisfies the normal equation.

$$A^t A\mathbf{x} = A^t \mathbf{b}.$$

If the columns of A are linearly independent then $Ker(A) = 0$ so A^tA is symmetric positive definite. Hence the solution of least square problem is unique. Assume this, and consider Cholesky decomposition which is stable.

$$LDL^t = A^tA.$$

This suggests the following approach to solving least square problem:

(1) Form $B = A^tA$ and the right hand side $\mathbf{c} = A^t\mathbf{b}$.

(2) Compute the Cholesky decomposition $B = LDL^t$.

(3) Solve $LDL^t\mathbf{x} = \mathbf{c}$ with forward and backward substitution.

Advantages of normal equation approach.

(1) The Cholesky decomposition LDL^t does not require partial pivoting for stability. Thus symmetry permutation may be used to lessen the fill-in during the decomposition.

(2) The computation of A^tA is carried out by summing along the columns of A so $b_{ij} = \sum_k a_{ki}a_{kj}$. Hence the row-reordering of A is irrelevant and B can be formed by processing the rows of A sequentially in any order.

(3) The decomposition $LDL^t = A^tA$ provides a convenient access to the useful statistical information contained in the unscaled covariance matrix $(A^tA)^{-1}$.

Disadvantages of normal equation approach.

(1) Unless extended precision is employed, there may be significant loss of information during the formation of A^tA.

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ \epsilon & 0 & 0 & 0 \\ 0 & \epsilon & 0 & 0 \\ 0 & 0 & \epsilon & 0 \\ 0 & 0 & 0 & \epsilon \end{pmatrix} \Rightarrow A^tA = \begin{pmatrix} 1+\epsilon^2 & 1 & 1 & 1 \\ 1 & 1+\epsilon^2 & 1 & 1 \\ 1 & 1 & 1+\epsilon^2 & 1 \\ 1 & 1 & 1 & 1+\epsilon^2 \end{pmatrix}$$

If $|\epsilon| < \sqrt{\text{machine number}} \approx 10^{-4}$, then the computed A^tA will be singular.

(2) The condition number of A^tA is quite large. The condition number of $m \times n$ matrix is defined as $\kappa_2(A) = \|A\|_2\|A^+\|_2$. In fact, the condition number of A^tA is $[\kappa_2(A)]^2$ which is $\frac{\sigma_1^2}{\sigma_r^2}$. Hence the normal equation produce am amplification of errors proportional to $[\kappa_2(A)]^2$.