Sistemas de Ecuaciones Lineales V

- **Métodos Iterativos:** Matrices dispersas. Esquema general. Métodos de Jacobi y de Gauss-Seidel.

Matrices dispersas

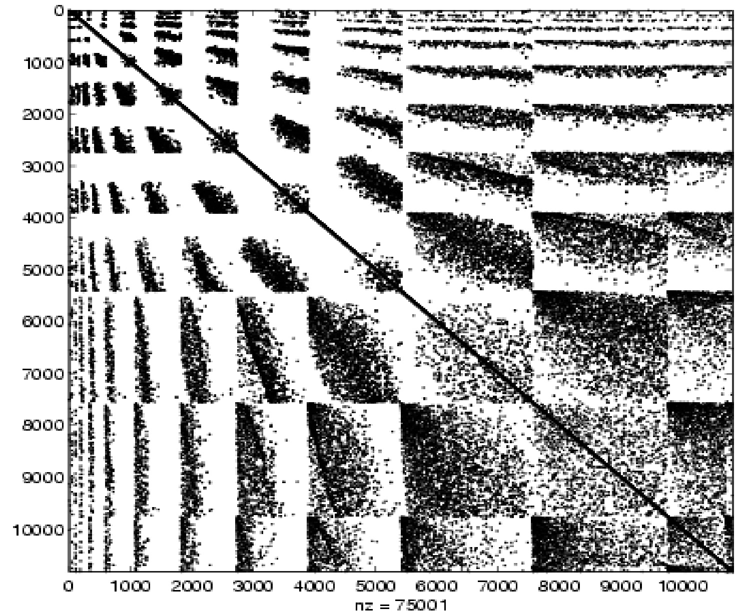
- Cuando la matriz A del sistema a resolver es dispersa, pero no banda, los métodos (directos) estudiados hasta ahora (eliminación de Gauss o Cholesky) presentan el defecto denominado **llenado** (**fill-in**).

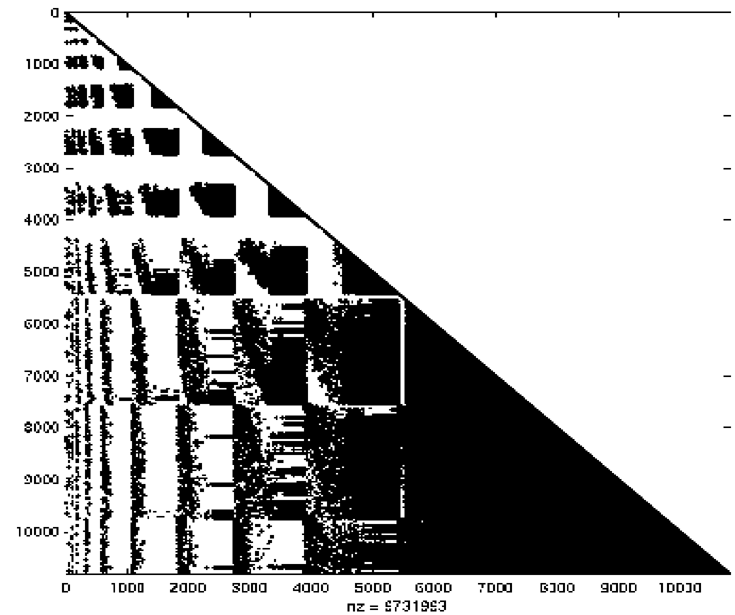
- El llenado consiste en que, a medida que el proceso de eliminación avanza, se van creando elementos no nulos en posiciones de L y U en donde la matriz A tiene ceros.

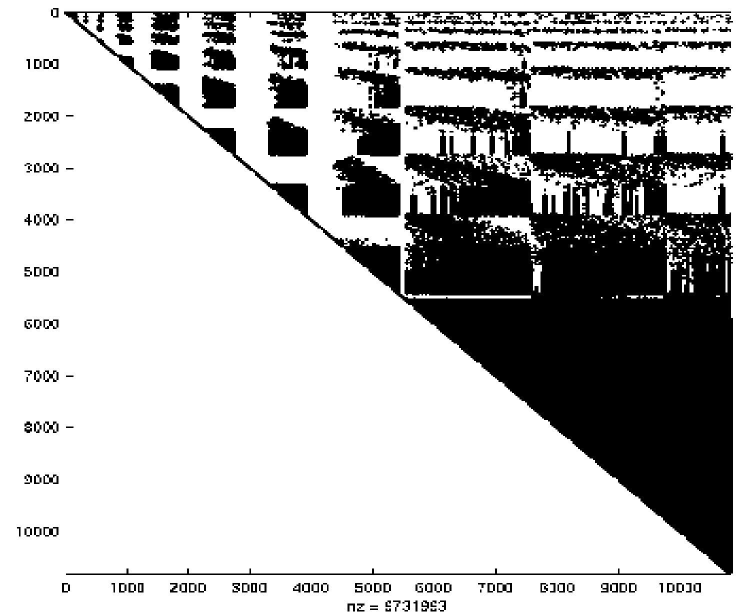
- Como consecuencia del llenado se tiene, por una parte, el aumento del número de flop y con ello el aumento del error de redondeo. Por otra parte se tiene el aumento en las necesidades de memoria para almacenar las matrices L y U, lo que puede llegar a hacer imposible aplicar estos métodos cuando A es de gran tamaño.

- Los métodos que estudieremos en seguida, llamados **iterativos**, evitan el llenado y sus consecuencias, al trabajar resolviendo reiteradamente sistemas con matriz diagonal o triangular-dispersa.

Llenado de matrices dispersas







Llenado de matrices dispersas (cont.)

```
>> load data.0125.mat
>> [L,U]=lu(A);
>> whos
  Name        Size                   Bytes  Class

  A       10821x10821               951580  sparse array
  L       10821x10821            151325524  sparse array
  U       10821x10821            151325524  sparse array
  b       10821x1                   129860  sparse array

Grand total is 25300219 elements using 303732496 bytes
```

Esquema general

- Considere el sistema de ecuaciones

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

con $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ no singular y $\boldsymbol{b} \in \mathbb{R}^n$.

Un **método iterativo** para resolver el sistema construye, a partir de un vector inicial $\boldsymbol{x}^{(0)}$, una sucesión de vectores $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(k)}, \ldots$ la que, bajo condiciones apropiadas, resultará convergente a $\boldsymbol{x}$.

- Si suponemos $\boldsymbol{A} = \boldsymbol{N} - \boldsymbol{P}$, donde $\boldsymbol{N}$ debe ser invertible, entonces

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \iff \boldsymbol{N}\boldsymbol{x} = \boldsymbol{P}\boldsymbol{x} + \boldsymbol{b} \iff \boldsymbol{x} = \boldsymbol{N}^{-1}\boldsymbol{P}\boldsymbol{x} + \boldsymbol{N}^{-1}\boldsymbol{b}.$$

Esquema general (cont.)

- Se usa la igualdad $\boldsymbol{N}\boldsymbol{x} = \boldsymbol{P}\boldsymbol{x} + \boldsymbol{b}$ para definir un esquema general para construir la sucesión $\{\boldsymbol{x}^{(k)}\}$.

- **Algoritmo del esquema general:**

> Dado el vector inicial $\boldsymbol{x}^{(0)}$,
>
> para $k = 1, 2, \ldots$ resolver:
>
> $$\boldsymbol{N}\boldsymbol{x}^{(k)} = \boldsymbol{P}\boldsymbol{x}^{(k-1)} + \boldsymbol{b},$$
>
> hasta que se satisfaga un criterio de detención.

- Definiendo $\boldsymbol{M} := \boldsymbol{N}^{-1}\boldsymbol{P}$ (**matriz de iteración**) y $\boldsymbol{e}^{(k)} := \boldsymbol{x} - \boldsymbol{x}^{(k)}$ (**error de $x^{(k)}$**), para cada $k = 1, 2, \ldots$ se tiene

$$\boldsymbol{e}^{(k)} = \boldsymbol{M}^k \boldsymbol{e}^{(0)}, \qquad k = 1, 2, \ldots$$

Convergencia de métodos iterativos.

- **Teorema.** (Convergencia) La sucesión $\left\{\boldsymbol{x}^{(k)}\right\}$ converge a la solución $\boldsymbol{x}$ de $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, si y sólo si, $\rho(\boldsymbol{M}) < 1$.

- **Observación.** Si la sucesión $\left\{\boldsymbol{x}^{(k)}\right\}$ converge, necesariamente lo hace a la solución $\boldsymbol{x}$ de $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

- **Lema.** (Cota para el radio espectral) Sea $\boldsymbol{A}$ una matriz cuadrada. Para cualquier norma matricial se tiene que

$$\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}\| .$$

- **Corolario.** (Condición suficiente de convergencia) Una condición suficiente para que la sucesión $\left\{\boldsymbol{x}^{(k)}\right\}$ sea convergente a la solución $\boldsymbol{x}$ de $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ es que

$$\|\boldsymbol{M}\| < 1,$$

donde $\boldsymbol{M}$ es la matriz de iteración del método que genera a $\left\{\boldsymbol{x}^{(k)}\right\}$.

Criterio de detención

- **Detención del proceso.** Cuando el proceso iterativo es convergente, éste se debe detener para un $\boldsymbol{x}^{(k+1)}$ tal que $\left\|\boldsymbol{e}^{(k+1)}\right\| = \left\|\boldsymbol{x} - \boldsymbol{x}^{(k+1)}\right\| \leq \texttt{tol}$, donde $\texttt{tol}$ indica un nivel de tolerancia prefijado para el error.

- **Lema.** Para $\|\boldsymbol{M}\| < 1$, se tiene que

$$\left\|\boldsymbol{x} - \boldsymbol{x}^{(k+1)}\right\| \leq \frac{\|\boldsymbol{M}\|}{1 - \|\boldsymbol{M}\|} \left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right\|.$$

- Un criterio de detención usual consiste en detener el proceso cuando

$$\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right\| \leq \texttt{tol}.$$

  Sin embargo, **este criterio resulta muchas veces inadecuado!**

  En efecto, si $\dfrac{\|\boldsymbol{M}\|}{1 - \|\boldsymbol{M}\|} \gg 1$, usualmente,

$$\frac{\|\boldsymbol{M}\|}{1 - \|\boldsymbol{M}\|} \left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right\| > \texttt{tol}.$$

Criterio de detención (cont.)

- **Observación.** Al graficar

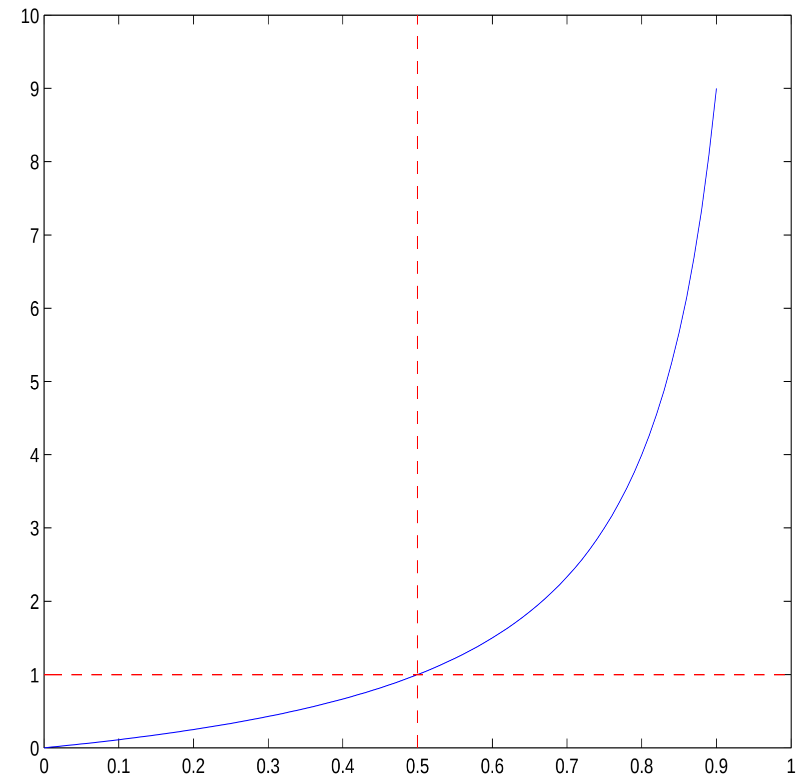
$$F(\boldsymbol{M}) := \frac{\|\boldsymbol{M}\|}{1 - \|\boldsymbol{M}\|},$$

se puede ver que:

$$\begin{aligned} \|\boldsymbol{M}\| \leq \frac{1}{2} \quad &\Longrightarrow \quad F(\boldsymbol{M}) \leq 1, \\ \|\boldsymbol{M}\| > \frac{1}{2} \quad &\Longrightarrow \quad F(\boldsymbol{M}) > 1. \end{aligned}$$

Luego, si $\|\boldsymbol{M}\| > \frac{1}{2}$, puede ser incorrecto detener el proceso cuando sólo se tiene

$$\left\| \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} \right\| \leq \texttt{tol}.$$



Criterio de detención (cont.)

- El criterio de detención implica calcular $\|\boldsymbol{M}\|$, lo que en general es difícil. El siguiente lema indica una manera de estimar $\|\boldsymbol{M}\|$.

- **Lema.** Para $k = 1, 2, \dots$ se tiene:

$$m_k := \frac{\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right\|}{\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right\|} \leq \|\boldsymbol{M}\|.$$

Demostración.

$$m_k = \frac{\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right\|}{\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right\|} = \frac{\left\|\left[\boldsymbol{x}^{(k+1)} - \boldsymbol{x}\right] - \left[\boldsymbol{x}^{(k)} - \boldsymbol{x}\right]\right\|}{\left\|\left[\boldsymbol{x}^{(k)} - \boldsymbol{x}\right] - \left[\boldsymbol{x}^{(k-1)} - \boldsymbol{x}\right]\right\|}$$

$$= \frac{\left\|\boldsymbol{e}^{(k+1)} - \boldsymbol{e}^{(k)}\right\|}{\left\|\boldsymbol{e}^{(k)} - \boldsymbol{e}^{(k-1)}\right\|} = \frac{\left\|\boldsymbol{M}\left[\boldsymbol{e}^{(k)} - \boldsymbol{e}^{(k-1)}\right]\right\|}{\left\|\boldsymbol{e}^{(k)} - \boldsymbol{e}^{(k-1)}\right\|} \leq \max_{\boldsymbol{y}\in\mathbb{R}^n:\,\boldsymbol{y}\neq 0} \frac{\|\boldsymbol{M}\boldsymbol{y}\|}{\|\boldsymbol{y}\|} = \|\boldsymbol{M}\|.$$

- En el criterio de detención puede utilizarse m_k como una estimación de $\|\boldsymbol{M}\|$.

  En tal caso, el proceso iterativo se detendrá cuando:

$$\frac{m_k}{1 - m_k}\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right\| \leq \mathtt{tol}.$$

Descomposición de una matriz

- Se considera resolver un sistema $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ con $a_{ii} \neq 0$, para $i = 1, \ldots, n$.

  Sea $\boldsymbol{x}^{(0)} = \left(x_1^{(0)}, \ldots, x_n^{(0)}\right)^{\mathrm{t}}$ arbitrario y escribamos la matriz $\boldsymbol{A}$ en la forma

$$\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{E} - \boldsymbol{F},$$

  donde $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{A})$, $-\boldsymbol{E}$ y $-\boldsymbol{F}$ son:

$$\boldsymbol{A} = \begin{pmatrix} \ddots & & -\boldsymbol{F} \\ & \boldsymbol{D} & \\ -\boldsymbol{E} & & \ddots \end{pmatrix}$$

- Notemos que tanto $\boldsymbol{D}$ como $\boldsymbol{D} - \boldsymbol{E}$ son matrices invertibles, ya que $a_{ii} \neq 0$ para $i = 1, \ldots, n$.

Método de Jacobi

- El **método de Jacobi** corresponde al esquema iterativo general con

$$\boldsymbol{N} := \boldsymbol{D} \qquad \text{y} \qquad \boldsymbol{P} := \boldsymbol{E} + \boldsymbol{F}.$$

- **Algoritmo de Jacobi:**

Dado el vector inicial $\boldsymbol{x}^{(0)}$,

para $k = 1, 2, \ldots$ resolver:

$$\boldsymbol{D}\boldsymbol{x}^{(k)} = (\boldsymbol{E} + \boldsymbol{F})\boldsymbol{x}^{(k-1)} + \boldsymbol{b},$$

hasta que se satisfaga un criterio de detención.

- En la iteración k, el vector $\boldsymbol{x}^{(k)}$ puede obtenerse **por componentes** como sigue:

Para $i = 1, \ldots, n$:

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k-1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right].$$

Método de Jacobi (cont.)

- La matriz de iteración del método de Jacobi verifica:

$$\boldsymbol{M} = \boldsymbol{D}^{-1}(\boldsymbol{E}+\boldsymbol{F}) = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 & -\frac{a_{n-1\,n}}{a_{n-1\,n-1}} \\ -\frac{a_{n1}}{a_{nn}} & \cdots & \cdots & -\frac{a_{n\,n-1}}{a_{nn}} & 0 \end{pmatrix}$$

- Para la norma infinito de $\boldsymbol{M}$ se tiene que $\|\boldsymbol{M}\|_\infty = \max_{1\leq i\leq n} \left\{ \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \right\}.$

- Cuando $\boldsymbol{A}$ es de **diagonal dominante estricta**, es decir, cuando se tiene que

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n,$$

entonces $\|\boldsymbol{M}\|_\infty < 1$ y el método de Jacobi resulta convergente.

Método de Gauss-Seidel

- El **método de Gauss-Seidel** corresponde al esquema iterativo general con

$$\boldsymbol{N} := \boldsymbol{D} - \boldsymbol{E} \qquad \text{y} \qquad \boldsymbol{P} := \boldsymbol{F}.$$

La matriz de iteración es entonces $\boldsymbol{M} = (\boldsymbol{D} - \boldsymbol{E})^{-1}\boldsymbol{F}$.

- **Algoritmo de Gauss-Seidel:**

> Dado el vector inicial $\boldsymbol{x}^{(0)}$,
>
> para $k = 1, 2, \ldots,$ resolver:
>
> $$(\boldsymbol{D} - \boldsymbol{E})\boldsymbol{x}^{(k)} = \boldsymbol{F}\boldsymbol{x}^{(k-1)} + \boldsymbol{b},$$
>
> hasta que se satisfaga un criterio de detención.

- En la iteración k, el vector $\boldsymbol{x}^{(k)}$ puede obtenerse **por componentes** como sigue:

> Para $i = 1, \ldots, n$:
>
> $$x_i^{(k)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right].$$

Notemos que esto corresponde a aprovechar en el paso k, los valores $x_j^{(k)}$ ya calculados.

Convergencia de los métodos de Jacobi y de Gauss-Seidel

- **Teorema.** Si A es de diagonal dominante estricta, entonces los métodos de Jacobi y de Gauss-Seidel convergen.

- **Observación.** Para una matriz arbitraria A, la convergencia de uno de estos métodos no implica la convergencia del otro.

- **Teorema.** Si A es simétrica y definida positiva, el método de Gauss-Seidel es convergente.

- **Observación.** Aunque A sea simétrica y definida positiva, el método de Jacobi puede ser divergente.

Convergencia de los métodos (cont.)

- **Ejemplo.** Para $s \in \mathbb{R}$, considere la matriz simétrica

$$\boldsymbol{A} = \begin{pmatrix} 1 & s & s \\ s & 1 & s \\ s & s & 1 \end{pmatrix},$$

  cuyos valores propios son: $1 - s$ (con multiplicidad 2) y $1 + 2s$.

  La matriz $\boldsymbol{A}$ es definida positiva cuando $s \in (-0.5, 1)$ y es de diagonal dominante estricta para $s \in (-0.5, 0.5)$.

- Se resolvió el sistema $\boldsymbol{Ax} = \boldsymbol{b}$ para un par de valores de s, usando los métodos de Jacobi y de Gauss-Seidel, con $\boldsymbol{b} = (1, 1, 1)^{\mathrm{t}}$ y $\boldsymbol{x}^{(0)} = (0.5, 0.5, 0.5)^{\mathrm{t}}$.

- Para $s = 0.3$, Jacobi itera 37 veces y Gauss-Seidel 12 veces (en ambos casos se implementó el criterio de detención visto en clase, con una tolerancia de 10^{-8}).

  Ambos métodos entregan como solución $\boldsymbol{x} = (0.6250, 0.6250, 0.6250)^{\mathrm{t}}$, que es la solución exacta.

Convergencia de los métodos (cont.)

- Para $s = 0.8$, en las mismas condiciones anteriores, Gauss-Seidel converge en la iteración 53 a

  $$(0.384\,615\,391\,735, 0.384\,615\,381\,035, 0.384\,615\,381\,784)^{\mathrm{t}}$$

  que difiere de la solución exacta

  $$\boldsymbol{x} = (0.384\,615\,384\,615, 0.384\,615\,384\,615, 0.384\,615\,384\,615)^{\mathrm{t}}$$

  en menos de $\texttt{tol} = 10^{-8}$ en cada componente.

- En cambio, al cabo de 100 iteraciones Jacobi entrega

  $$10^{19} \times (-1.862\,199\,431\,313, -1.862\,199\,431\,313, -1.862\,199\,431\,313)^{\mathrm{t}},$$

  vector que no tiene ninguna relación con la solución del sistema.

  Se nota claramente que, en este caso, el método de Jacobi diverge.