(Full) Newton Method

- Alternatively, we can write:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \delta\mathbf{x}^{(k)},$$

where the vector $\delta\mathbf{x}^{(k)}$ is referred to as the **Newton step**, or **Newton correction** and is the solution of the linear system:

$$\mathbf{J}(\mathbf{x}^{(k)})\delta\mathbf{x}^{(k)} = -\mathbf{F}(\mathbf{x}^{(k)}).$$

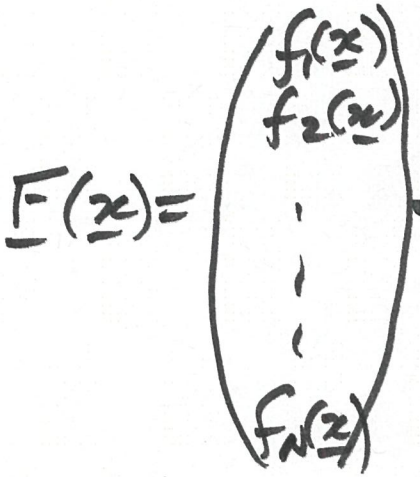
(Full) Newton Method

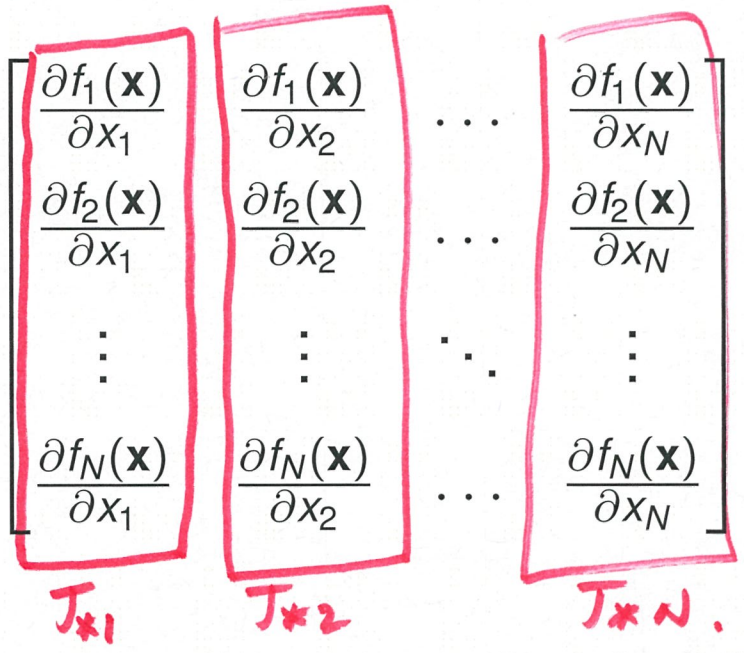
- The computational framework presented above can be extended to higher dimensions. Given an initial iterate $\mathbf{x}^{(0)}$, Newton's method for solving (1) takes the form:

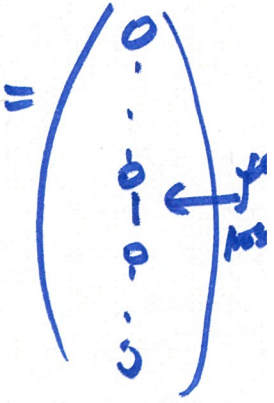
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{J}\left(\mathbf{x}^{(k)}\right)^{-1} \mathbf{F}\left(\mathbf{x}^{(k)}\right), \tag{6}$$

where $\mathbf{J} \in \mathbb{R}^{N \times N}$ is the Jacobian matrix of $\mathbf{F}$:

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} \nabla f_1(\mathbf{x})^T \\ \nabla f_2(\mathbf{x})^T \\ \vdots \\ \nabla f_N(\mathbf{x})^T \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_N} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_2(\mathbf{x})}{\partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_N(\mathbf{x})}{\partial x_1} & \frac{\partial f_N(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_N(\mathbf{x})}{\partial x_N} \end{bmatrix}$$







Terminating Newton's Method

- One common criterion used to terminate Newton's method is based on the norm of the nonlinear residual $\|\mathbf{F}(\mathbf{x}^{(k+1)})\|$.
- Typically, a combination of both a **relative error tolerance** τ_r and **absolute error tolerance** τ_a is used to halt the iteration:

$$\underbrace{\left\|\mathbf{F}\left(\mathbf{x}^{(k+1)}\right)\right\|}_{\text{nonlinear residual}} \leq \tau_r \left\|\mathbf{F}\left(\mathbf{x}^{(0)}\right)\right\| + \tau_a. \tag{7}$$

(Full) Newton Method

n^2 derivatives

An algorithm for Newton's method is summarised as follows.

Newton's method

Choose $\mathbf{x}^{(0)}$

While not converged for $k = 0, 1, \ldots$

Solve $\mathbf{J}(\mathbf{x}^{(k)})\delta\mathbf{x}^{(k)} = -\mathbf{F}(\mathbf{x}^{(k)})$ for $\delta\mathbf{x}^{(k)}$

$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \delta\mathbf{x}^{(k)}$

End

$\|E(\underline{x}^{(k+1)})\|_2 < \varepsilon$

Generate J: n^2 function evaluations

PLU decomposition.

$$\|\underline{x}\|_p = \left[\sum_{i=1}^{N} |x_i|^p\right]^{1/p},$$

$p = 1$: $\|\underline{x}\|_1 = \sum_{i=1}^{N} |x_i|$

$p = 2$: $\|x\|_2 = \sqrt{\sum_{i=1}^{N} |x_i|^2}$

$p = \infty$: $\|\underline{x}\|_\infty = \max_{1 \le i \le N} |x_i|$.

Chord Method

An algorithm for the Chord method is summarised as follows.

Chord method

Choose $\mathbf{x}^{(0)}$ and set $\bar{\mathbf{J}} = \mathbf{J}(\mathbf{x}^{(0)})$

While not converged for $k = 0, 1, \ldots$

Solve $\bar{\mathbf{J}}\delta\mathbf{x}^{(k)} = -\mathbf{F}(\mathbf{x}^{(k)})$ for $\delta\mathbf{x}^{(k)}$

$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \delta\mathbf{x}^{(k)}$

End

Shamanskii Method

Alternatively, the Jacobian matrix may be *periodically* updated. For example, alternating a full Newton step with a sequence of Chord steps leads to a method often attributed to Shamanskii.

tola, tolr, $\rho^* = 0.5$

Shamanskii method

Choose $\mathbf{x}^{(0)}$

err = $\|F^{(0)}\|$
tol = tola + tolr * err .

While not converged for $k = 0, 1, \ldots$

If $k \equiv 0 \ (mod\ m)$ or $\rho > \rho^*$

Compute/Update $\bar{\mathbf{J}} = \mathbf{J}(\mathbf{x}^{(k)})$

End

Solve $\bar{\mathbf{J}}\delta\mathbf{x}^{(k)} = -\mathbf{F}(\mathbf{x}^{(k)})$ for $\delta\mathbf{x}^{(k)}$

$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \delta\mathbf{x}^{(k)}$

errold = err ;
err = $\|F^{(k+1)}\|$
ρ = err / errold ;

End

ρ is the ratio of nonlinear residuals

$\frac{err}{errold} < \rho^* = 0.5$

$err < 0.5\, errold$

Convergence Analysis

Definition (Lipschitz continuity)

Let $D \subset \mathbb{R}^N$ and $\mathbf{J} : D \to \mathbb{R}^{N \times N}$. Then $\mathbf{J}$ is **Lipschitz continuous** on D, with **Lipschitz constant** γ if

$$\|\mathbf{J}(\mathbf{x}) - \mathbf{J}(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|, \ \forall \mathbf{x}, \mathbf{y} \in D,$$

and we write $\mathbf{J} \in \mathrm{Lip}_\gamma(D)$.

The Standard Assumptions

1. $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ has a solution $\mathbf{x}^*$ on $D \subset \mathbb{R}^N$.
2. $\mathbf{J} \in \mathrm{Lip}_\gamma(D)$.
3. $\mathbf{J}(\mathbf{x})$ is nonsingular and $\|\mathbf{J}(\mathbf{x})^{-1}\| \leq \beta \ \ \forall \mathbf{x} \in D$.

Convergence Analysis

The following formulation of the fundamental theorem of calculus will be important in the theory that follows.

Theorem

Let $\mathbf{F} : D \to \mathbb{R}^N$ *be differentiable on an open subset* $D \subset \mathbb{R}^N$ *and let* $\mathbf{x} \in D$*. Then for any nonzero vector* $\mathbf{h}$*, such that* $\mathbf{x} + \mathbf{h} \in D$*, we have*

$$\mathbf{F}(\mathbf{x} + \mathbf{h}) = \mathbf{F}(\mathbf{x}) + \int_0^1 \mathbf{J}(\mathbf{x} + t\mathbf{h})\,\mathbf{h}\,dt$$

where $\mathbf{J}$ *is the Jacobian matrix of* $\mathbf{F}$*.*

Finite Difference Jacobian

- If $\mathbf{F}$ has a complicated form, or if it is only available as a computer code rather than an analytic formula, it may not be possible to compute the Jacobian matrix analytically. In this case finite difference approximations to the derivatives can be used instead.
- Typically a first order forward difference approximation is used, whereby the j^{th} column of the Jacobian matrix is approximated as follows:

$$\mathbf{J}_{*,j}(\mathbf{x}) = \mathbf{J}(\mathbf{x})\mathbf{e}_j \approx \widetilde{\mathbf{J}}_{*,j}(\mathbf{x}) = \frac{\mathbf{F}(\mathbf{x} + h\mathbf{e}_j) - \mathbf{F}(\mathbf{x})}{h} \tag{9}$$

  for a suitable shift value $h > 0$, where $\mathbf{e}_j$ is the j^{th} column of the $N \times N$ identity matrix.
- This strategy can be applied to either Newton's method, or the Chord and Shamanskii methods.

Idea: choose $\underline{h} = \varepsilon \underline{e}_j$, $\underline{e}_j = (0, \ldots, 0, 1, 0 \ldots 0)^T$.
ε is a small number.

$$\underline{F}(\underline{x} + \varepsilon \underline{e}_j) = \underline{F}(\underline{x}) + \int_0^1 \varepsilon J(\underline{x} + t\varepsilon \underline{e}_j)\underline{e}_j \, dt .$$

$$J_{*j}(\underline{x}) = J(\underline{x})\underline{e}_j \approx \frac{\underline{F}(\underline{x} + \varepsilon \underline{e}_j) - \underline{F}(\underline{x})}{\varepsilon} = \tilde{J}_{*j}(\underline{x})$$

What is the error in this approximation?
First error in j^{th} column:

$$\| J_{*j}(\underline{x}) - \tilde{J}_{*j}(\underline{x}) \| \leq C\varepsilon \; ?$$

$$J_{*j}(\underline{x}) - \tilde{J}_{*j}(x) = J(\underline{x})\underline{e}_j - \left\{ \frac{\underline{F}(\underline{x} + \varepsilon \underline{e}_j) - \underline{F}(\underline{x})}{\varepsilon} \right\}$$

$$= J(\underline{x})\underline{e}_j - \int_0^1 J(\underline{x} + t\varepsilon \underline{e}_j)\, \underline{e}_j \, dt$$

$$= \int_0^1 J(\underline{x})\underline{e}_j \, dt - \int_0^1 J(\underline{x} + t\varepsilon \underline{e}_j)\underline{e}_j \, dt$$

$$\|J_{*j} - \tilde{J}_{*j}\| = \left\| \int_0^1 \left[J(\underline{x}) - J(\underline{x} + t\varepsilon \underline{e}_j) \right] \underline{e}_j \, dt \right\|$$

$\|A\underline{x}\| \leq \|A\| \|\underline{x}\|$

(using matrix calculus + vector/matrix norm theory)

$$\leq \int_0^1 \|(J(\underline{x}) - J(\underline{x} + t\varepsilon \underline{e}_j))\underline{e}_j\| \, dt$$

$$\leq \int_0^1 \|J(\underline{x}) - J(\underline{x} + t\varepsilon \underline{e}_j)\| \, dt$$

matrix-vector norm compatibility + $\|\underline{e}_j\| = 1$.

$$\leq \int_0^1 \gamma \|\underline{x} - \underline{x} - t\varepsilon \underline{e}_j\| \, dt$$

Assume that $J \in Lip_\gamma(D)$

$$\leq \varepsilon \int_0^1 \gamma |t| \, dt = \frac{\varepsilon \gamma}{2}$$

$\|\alpha \underline{x}\| = |\alpha| \|\underline{x}\|$.

$$\|J - \tilde{J}\|_2^2 \leq \|J - \tilde{J}\|_F^2 = \sum_{j=1}^{n} \|(J - \tilde{J})\underline{e}_j\|_2^2$$

$$\leq n \frac{\varepsilon^2 \gamma^2}{2^2} =$$

$$\boxed{\|J - \tilde{J}\|_2 \leq \sqrt{n} \, \frac{\varepsilon \gamma}{2}}$$

$$J(\underline{x}) = \begin{pmatrix} \frac{\partial f_1(\underline{x})}{\partial x_1} & \frac{\partial f_1(\underline{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\underline{x})}{\partial x_j} & \cdots & \frac{\partial f_1(\underline{x})}{\partial x_N} \\ \frac{\partial f_2(\underline{x})}{\partial x_1} & \frac{\partial f_2(\underline{x})}{\partial x_2} & & \frac{\partial f_2(\underline{x})}{\partial x_j} & & \frac{\partial f_2(\underline{x})}{\partial x_N} \\ \vdots & \vdots & & \vdots & & \vdots \\ \frac{\partial f_N(\underline{x})}{\partial x_1} & \frac{\partial f_N(\underline{x})}{\partial x_2} & & \frac{\partial f_N}{\partial x_j}(\underline{x}) & \cdots & \frac{\partial f_N}{\partial x_N}(\underline{x}) \end{pmatrix}$$

$\underline{F}(\underline{x}) = \underline{F}$

$\frac{\underline{F}(\underline{x}+\varepsilon\underline{e}_1)-\underline{F}}{\varepsilon}$ $\quad \frac{\underline{F}(\underline{x}+\varepsilon\underline{e}_2)-\underline{F}}{\varepsilon}$ $\quad \frac{\underline{F}(\underline{x}+\varepsilon\underline{e}_j)-\underline{F}}{\varepsilon}$ $\quad (\underline{F}(\underline{x}+\varepsilon\underline{e}_N)-\underline{F})/\varepsilon$

Finite Difference Jacobian

Theorem

Let $\mathbf{F} : D \to \mathbb{R}^N$ *for some open subset* $D \subset \mathbb{R}^N$ *and let* $\mathbf{J} \in Lip_\gamma(D)$, *then*

$$\|\mathbf{J}_{*,j}(\mathbf{x}) - \widetilde{\mathbf{J}}_{*,j}(\mathbf{x})\| \leq \frac{\gamma}{2} h$$

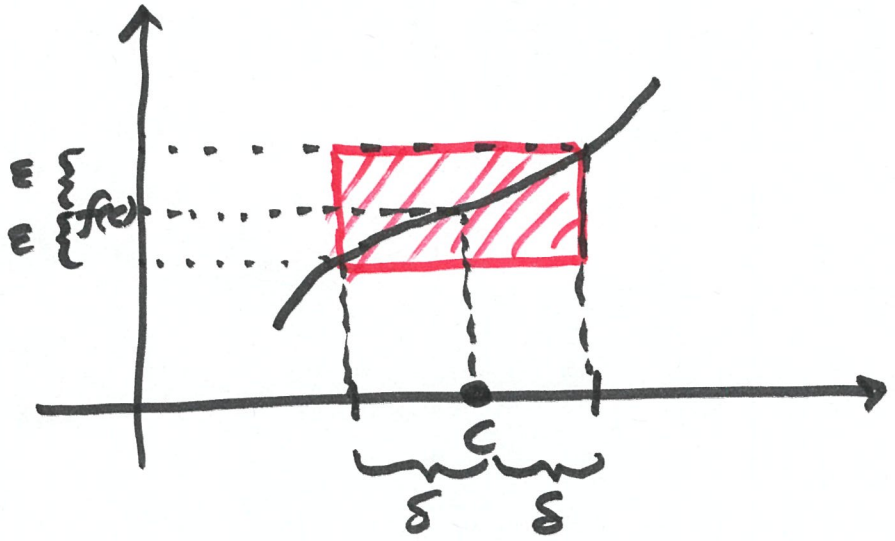
where $\widetilde{\mathbf{J}}_{*,j}$ *is computed using finite differences as described in equation (9).*

One reasonable choice for the shift parameter is

$$h = \begin{cases} \sqrt{\epsilon}\|\mathbf{x}\|_2 & \mathbf{x} \neq 0 \\ \sqrt{\epsilon} & \mathbf{x} = 0\,, \end{cases}$$

where ϵ is the unit roundoff (denoted by *eps* and equal to *2.2204e-16* in MATLAB).

Definition: Let $D \subset \mathbb{R}$ and $f: D \to \mathbb{R}$ then f is continuous at c if for every $\varepsilon > 0$ there is a $\delta > 0$ such that whenever $x \in D$ and $|x - c| < \delta$ then $|f(x) - f(c)| < \varepsilon$



for $|x - c| < \delta$ $f(x)$ must be in the shaded region

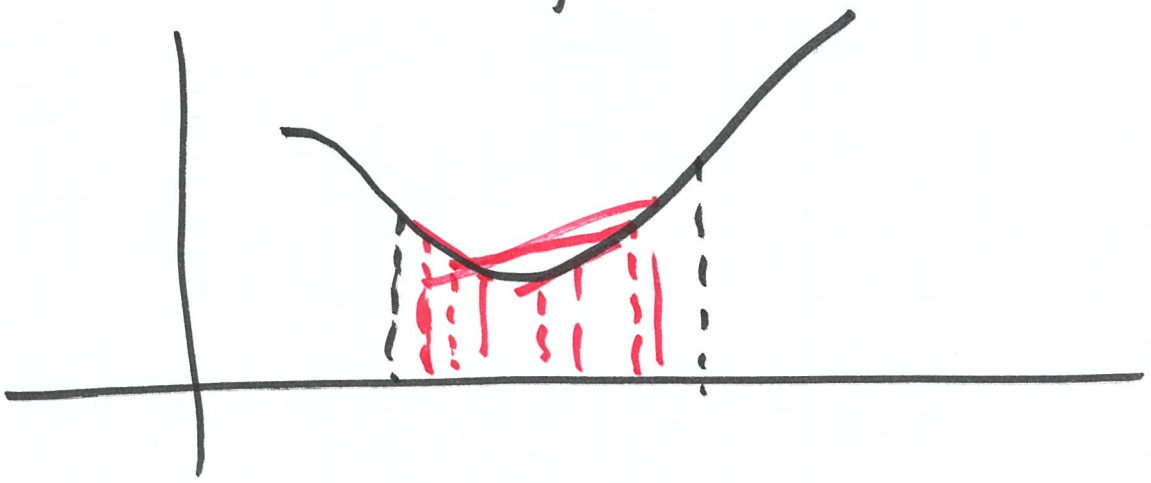
Note: δ depends on ε and c.

Lipschitz continuity:

$f: D \to \mathbb{R}$ such that there exists a number γ such that for all $x, y \in D$

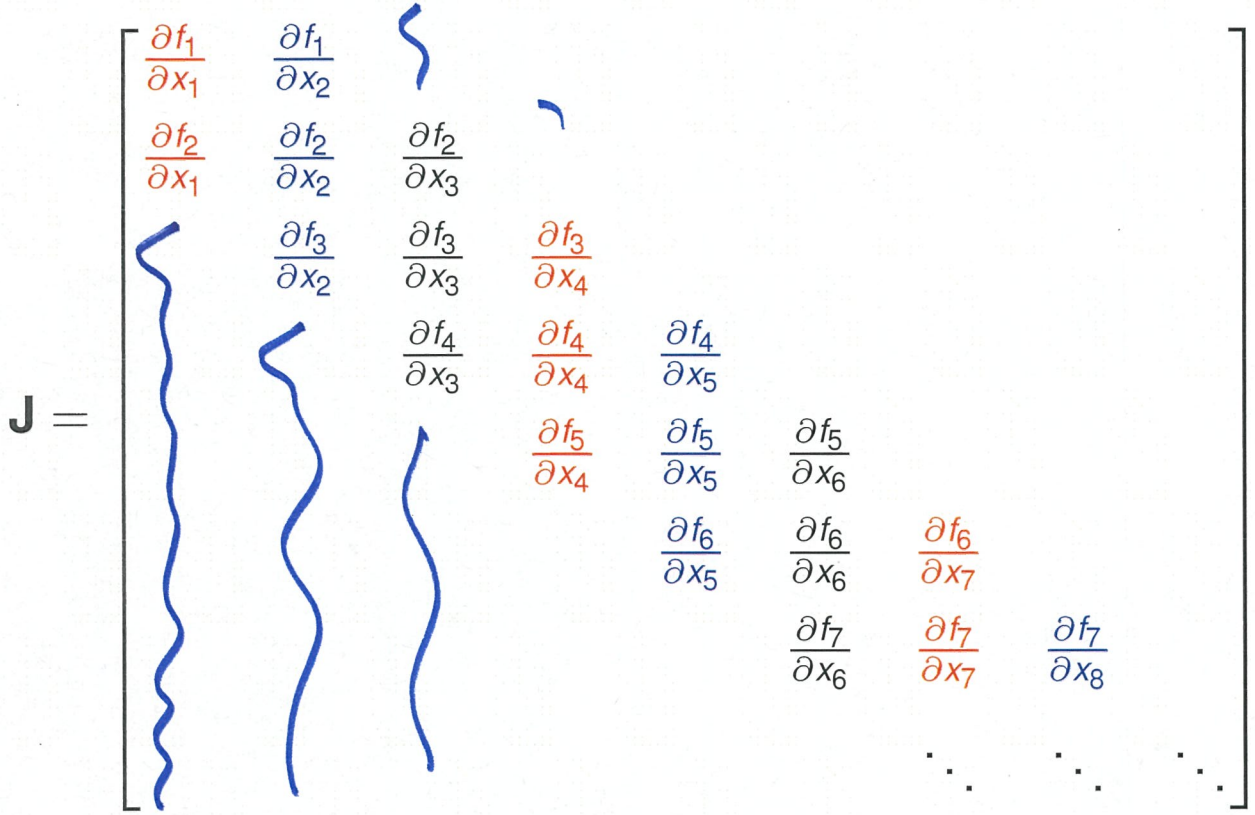
$$|f(x) - f(y)| \leq \gamma |x - y|$$

Interpretation: $\left| \frac{f(x) - f(y)}{x - y} \right| \leq \gamma$ slope of secant lines.



Tridiagonal Jacobian

Consider a tridiagonal Jacobian:



We now show that a Jacobian matrix having this structure can be approximated using only 3 additional function evaluations.

Tridiagonal Jacobian

Define the vectors in $\mathbb{R}^N$:

$$\mathbf{s}_1 = (1,0,0,1,0,0,1,\ldots)^T, \quad \mathbf{s}_2 = (0,1,0,0,1,0,0,\ldots)^T,$$
$$\mathbf{s}_3 = (0,0,1,0,0,1,0,\ldots)^T,$$

and consider the products $\mathbf{J}\mathbf{s}_1$, $\mathbf{J}\mathbf{s}_2$ and $\mathbf{J}\mathbf{s}_3$

- $\mathbf{J}\mathbf{s}_1 = \mathbf{J}_{*,1} + \mathbf{J}_{*,4} + \mathbf{J}_{*,7} + \ldots = \sum_{j=1:3:\mathtt{N}} \mathbf{J}_{*,j}$ is the sum of columns 1, 4, 7, . . . , and so on.
- $\mathbf{J}\mathbf{s}_2 = \mathbf{J}_{*,2} + \mathbf{J}_{*,5} + \mathbf{J}_{*,8} + \ldots = \sum_{j=2:3:\mathtt{N}} \mathbf{J}_{*,j}$ is the sum of columns 2, 5, 8, . . . , and so on.
- $\mathbf{J}\mathbf{s}_3 = \mathbf{J}_{*,3} + \mathbf{J}_{*,6} + \mathbf{J}_{*,9} + \ldots = \sum_{j=3:3:\mathtt{N}} \mathbf{J}_{*,j}$ is the sum of columns 3, 6, 9, . . . , and so on.

where for example `1:3:N` refers to MATLAB's colon notation.

Tridiagonal Jacobian

The key observation is that each of the above summations involve vectors that do not possess entries in the same row. Therefore:

$$\mathbf{Js}_1 = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} \\ \frac{\partial f_2}{\partial x_1} \\ \frac{\partial f_3}{\partial x_4} \\ \frac{\partial f_4}{\partial x_4} \\ \frac{\partial f_5}{\partial x_4} \\ \frac{\partial f_6}{\partial x_7} \\ \frac{\partial f_7}{\partial x_7} \\ \vdots \end{bmatrix}, \quad \mathbf{Js}_2 = \begin{bmatrix} \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_2} \\ \frac{\partial f_3}{\partial x_2} \\ \frac{\partial f_4}{\partial x_5} \\ \frac{\partial f_5}{\partial x_5} \\ \frac{\partial f_6}{\partial x_5} \\ \frac{\partial f_7}{\partial x_8} \\ \vdots \end{bmatrix}, \quad \mathbf{Js}_3 = \begin{bmatrix} \frac{\partial f_1}{\partial x_3} \\ \frac{\partial f_2}{\partial x_3} \\ \frac{\partial f_3}{\partial x_3} \\ \frac{\partial f_4}{\partial x_6} \\ \frac{\partial f_5}{\partial x_6} \\ \frac{\partial f_6}{\partial x_6} \\ \frac{\partial f_7}{\partial x_9} \\ \vdots \end{bmatrix}.$$

Tridiagonal Jacobian

For a given vector $\mathbf{x} \in \mathbb{R}^N$, each of the above vectors can be approximated as follows:

$$\mathbf{J}(\mathbf{x})\mathbf{s}_j \approx \frac{\mathbf{F}(\mathbf{x} + h_j\mathbf{s}_j) - \mathbf{F}(\mathbf{x})}{h_j} \tag{10}$$

where the shift parameter is chosen as

$$h_j = \begin{cases} \dfrac{\sqrt{\epsilon}\|\mathbf{x}\|_2}{\|\mathbf{s}_j\|_2} & \mathbf{x} \neq 0 \\ \dfrac{\sqrt{\epsilon}}{\|\mathbf{s}_j\|_2} & \mathbf{x} = 0\,. \end{cases}$$

The above strategy can also be extended to a general banded matrix with bandwidth greater than 3.