Applied Optimization: Formulation and Algorithms for Engineering Systems Slides

Ross Baldick

Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712

Copyright © 2018 Ross Baldick

Part II

Non-linear simultaneous equations

6

Case studies of non-linear simultaneous equations

(i) Solution of Kirchhoff's laws in a non-linear direct current (DC) circuit (Section 6.1), and

(ii) Solution of Kirchhoff's laws in a linear alternating current (AC) circuit where the variables of interest are not currents and voltages but instead are power (and "reactive power") injections (Section 6.2).

6.1 Analysis of a non-linear direct current circuit

6.1.1 Motivation

- Predict the behavior of the circuit without actually building a prototype.
- Predict the effect of changes in component values on the circuit behavior.

6.1.2 Formulation

6.1.2.1 Device models

Terminal characteristics

- For a resistor, the current is a linear function of voltage.

Non-linear devices

- Diode model:

$$\forall V_{\text{diode}} \in \mathbb{R}, i_{\text{diode}}(V_{\text{diode}}) = I_{\text{sat}} \left[\exp\left(\frac{qV_{\text{diode}}}{\eta \, \mathcal{K} T} \right) - 1 \right]. \tag{6.1}$$
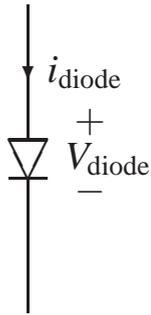

Fig. 6.1. Symbol for diode together with voltage and current conventions.

Non-linear devices, continued

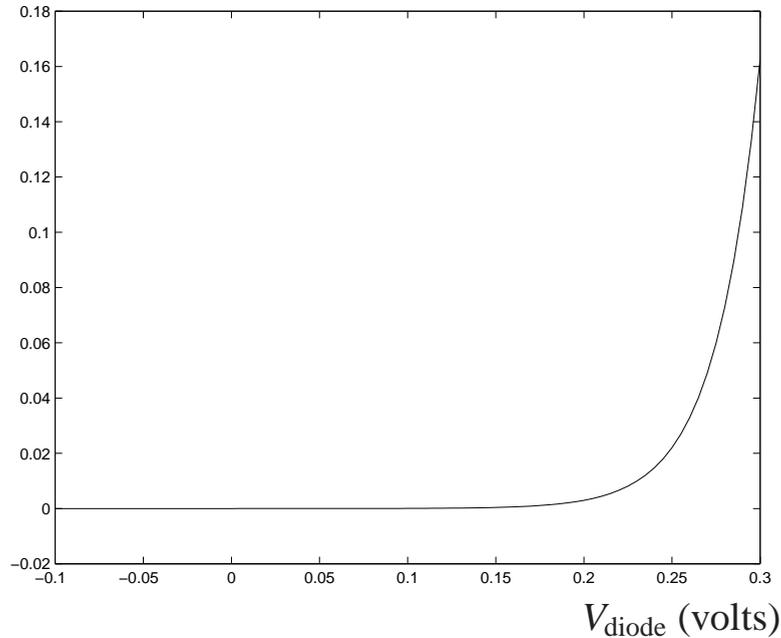$i_{\text{diode}}(V_{\text{diode}})$ (amps)

Fig. 6.2. Current to voltage relationship for diode.

Choice of terminal model

- Again, Occam's razor is important in selecting a terminal model.
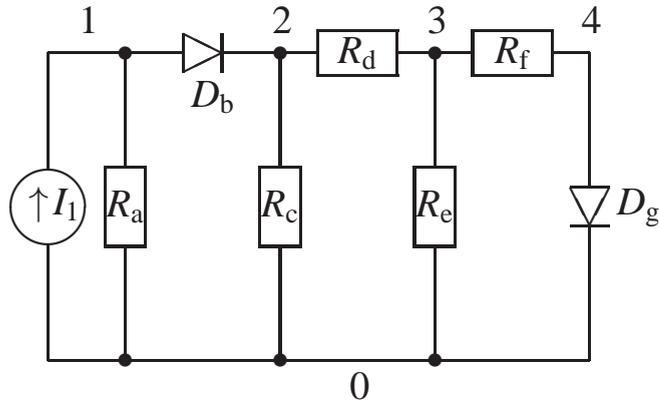
6.1.2.2 Kirchhoff's current law

Fig. 6.3. A simple non-linear circuit.

- By Kirchhoff's current law applied to nodes 1, 2, 3, 4:

$$\left(\frac{1}{R_a}\right) x_1 + i_b(x_1 - x_2) - I_1 = 0, \qquad (6.2)$$

$$-i_b(x_1 - x_2) + \left(\frac{1}{R_c} + \frac{1}{R_d}\right) x_2 + \left(-\frac{1}{R_d}\right) x_3 = 0, \qquad (6.3)$$

$$\left(-\frac{1}{R_d}\right) x_2 + \left(\frac{1}{R_d} + \frac{1}{R_e} + \frac{1}{R_f}\right) x_3 + \left(-\frac{1}{R_f}\right) x_4 = 0, \qquad (6.4)$$

$$\left(-\frac{1}{R_f}\right) x_3 + \left(\frac{1}{R_f}\right) x_4 + i_g(x_4) = 0. \qquad (6.5)$$

- As in the direct current linear circuit case study in Section 4.1, the equation for the datum node is **redundant**.

6.1.2.3 Non-linear equations

- Define the **vector function** $g : \mathbb{R}^4 \to \mathbb{R}^4$ by:

$$\forall x \in \mathbb{R}^4, g(x) = \begin{bmatrix} \left(\frac{1}{R_a}\right) x_1 + i_b(x_1 - x_2) - I_1 \\ -i_b(x_1 - x_2) + \left(\frac{1}{R_c} + \frac{1}{R_d}\right) x_2 + \left(-\frac{1}{R_d}\right) x_3 \\ \left(-\frac{1}{R_d}\right) x_2 + \left(\frac{1}{R_d} + \frac{1}{R_e} + \frac{1}{R_f}\right) x_3 + \left(-\frac{1}{R_f}\right) x_4 \\ \left(-\frac{1}{R_f}\right) x_3 + \left(\frac{1}{R_f}\right) x_4 + i_g(x_4) \end{bmatrix}. \quad (6.6)$$

- If we write $g(x) = \mathbf{0}$ then we have reproduced (6.2)–(6.5).
- These are a set of non-linear simultaneous equations.
- To represent linear equations $Ax = b$ in this way we would define $g : \mathbb{R}^n \to \mathbb{R}^m$ by:

$$\forall x \in \mathbb{R}^n, g(x) = Ax - b.$$

6.1.3 Circuit changes

- Changes in the values of resistors, current sources, or diode parameters will change the functional form of corresponding entries in g.
- For example, if a resistor or a diode between nodes ℓ and k changes then the functional form of g_ℓ and g_k will change.
- If a resistor, current source, or diode between node ℓ and the datum node changes then the functional form of g_ℓ will change.
- Changes in the diode could be due to changes in I_{sat}, η, or T, for example, and would change the functional relationship between the diode current and diode voltage.

6.1.4 Problem characteristics

6.1.4.1 Numbers of variables and equations

- As in the linear circuit, we have the same number of variables as equations.

6.1.4.2 Number of solutions

- The current to voltage characteristic of a diode is strictly monotonically increasing so that increasing voltage corresponds to increasing current.
- Strict monotonicity of component model functions is sufficient to guarantee that there is at most one solution for the circuit.
- Not every two-terminal electronic component has a strictly monotonically increasing terminal model.

- For example, a **tunnel diode** has a characteristic that is not strictly monotonically increasing.
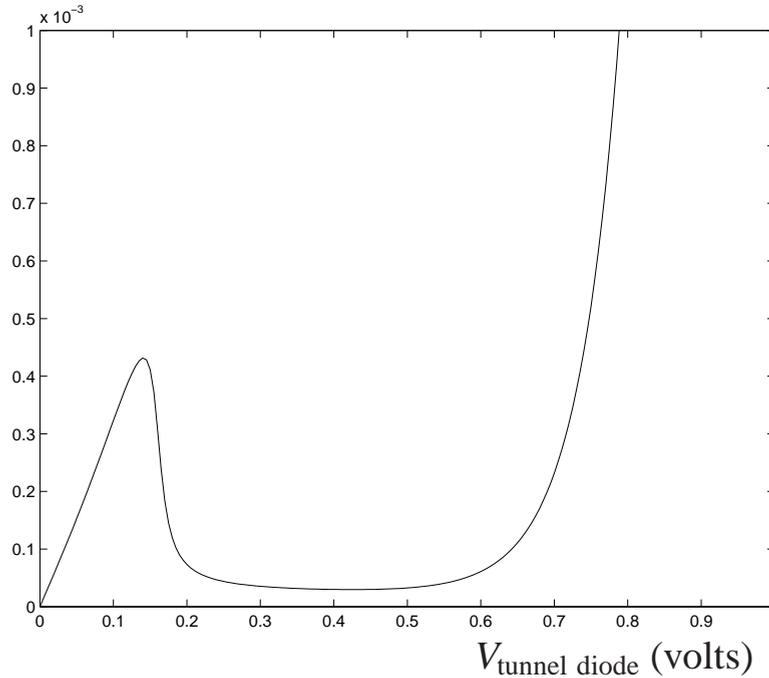
$I_{\text{tunnel diode}}(V_{\text{tunnel diode}})$ (amps)

$V_{\text{tunnel diode}}$ (volts)

Fig. 6.4. Current to voltage relationship for tunnel diode.

6.1.4.3 Sparsity

- The vector function g is "sparse" in the sense that a typical entry of g depends only on a few entries of x.
- Although we cannot store the representation of a non-linear function as a sparse matrix, we can still store the parameters necessary to specify the functions in a sparse structure as suggested in Figure 6.5.
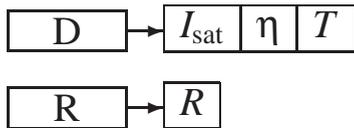

Fig. 6.5. Storage of parameters for diode and resistor as linked lists.

6.1.4.4 Non-existence of direct algorithms

- Because of the non-linear diode elements, there is in general no direct algorithm for solving an arbitrary circuit consisting of current sources, resistors, and diodes.

6.2 Analysis of an electric power system

6.2.1 Motivation

- It is important to be able to predict the power flows on lines and the voltage magnitudes at loads in advance of actual operations.

6.2.2 Formulation
6.2.2.1 Variables

Phasors

- We can use complex numbers, called **phasors,** to represent the magnitude and angle of the AC voltages and currents at a fixed frequency.
- The **magnitude** of the complex number represents the root-mean-square magnitude of the voltage or current.
- The **angle** of the complex number represents the angular displacement between the sinusoidal voltage or current and a reference sinusoid.

Reference angle

- The angles of the voltages and currents in the system would all change if we changed the angle of our reference sinusoid, but this would have no effect on the physical system.
- We can therefore arbitrarily assign the angle at one of the buses to be zero and measure all the other angles with respect to this angle.
- We call this bus the **reference bus**.

Representation of complex numbers

- To represent a complex number V with real numbers requires two real numbers, either:
  - the **magnitude** $|V|$ and the **angle** $\angle V$, so that $V = |V| \exp(\angle V \sqrt{-1})$, or
  - the **real** $\Re\{V\}$ and **imaginary** $\Im\{V\}$ parts, so that $V = \Re\{V\} + \Im\{V\} \sqrt{-1}$.

- Since we need to compare voltage magnitudes to limits, we will represent voltages as magnitudes and angles.

Scaling and "per unit"

- There are voltage transformers throughout a typical power system.
- This means that the nominal voltage magnitude varies considerably throughout the system by several orders of magnitude.
- We scale the voltage magnitude so that an actual value of 121 kV in the 110 kV part of the system would be represented by a scaled value of:

$$\frac{121 \text{ kV}}{110 \text{ kV}} = 1.1,$$

- while an actual value of 688.5 kV in the 765 kV part of the system would be represented by a scaled value of:

$$\frac{688.5 \text{ kV}}{765 \text{ kV}} = 0.9.$$

Three-phase circuits

- Generation-transmission systems are usually operated as balanced **three-phase systems**.
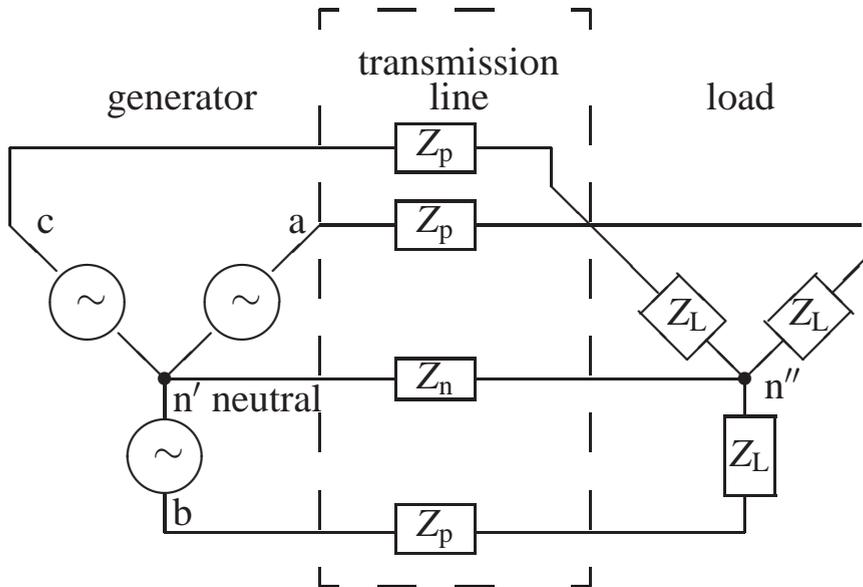

Fig. 6.6. An example balanced three-phase system.

Per-phase equivalent

- The behavior of a balanced three-phase circuit can be completely determined from the behavior of a **per-phase equivalent circuit**.
- Figure 6.7 shows the a-phase equivalent circuit of the three-phase circuit of Figure 6.6.
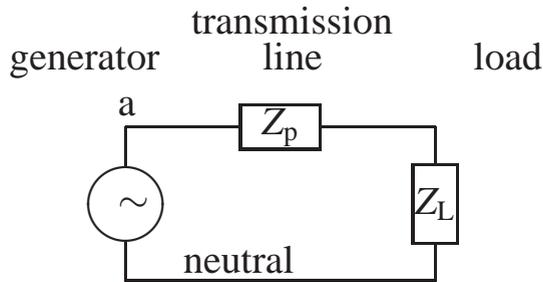

Fig. 6.7. Per-phase equivalent circuit for the three-phase circuit in Figure 6.6.

Model transformation

- The determination of the behavior of a three-phase system through the analysis of a related per-phase equivalent is an example of **model transformation** that utilizes the **symmetry** of the three-phase circuit.

6.2.2.3 Transmission lines

- We can represent the terminal behavior of **distributed parameter circuits** with a π-**equivalent** circuit.

- Each component has an impedance (or, equivalently, an **admittance**) determined by the characteristics of the line.
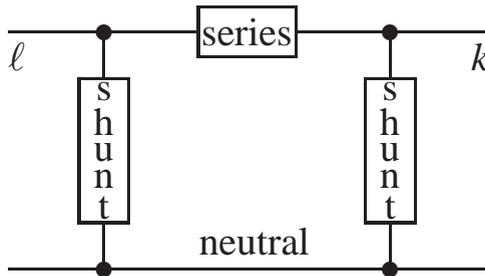
Fig. 6.8. Equivalent π circuit of per-phase equivalent of transmission line.

6.2.2.4 Bus admittance matrix and power flow equations

- Consider the per-phase equivalent of a three bus, three line transmission system as illustrated in Figure 6.9.
- For each bus $\ell = 1, 2, 3$, the pair of shunt π elements joining node ℓ to neutral can be combined together to form a single shunt element.
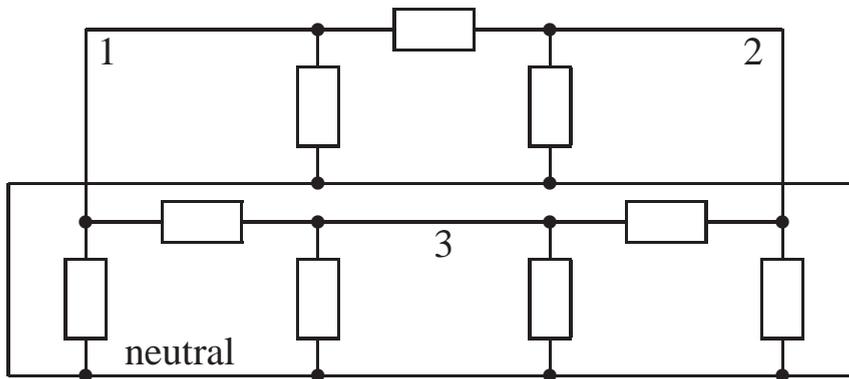
Fig. 6.9. Per-phase equivalent circuit model for three bus, three line system.

Bus admittance matrix and power flow equations, continued

- This yields a circuit with:
  - one element corresponding to each of the buses $\ell = 1, 2, 3$, joining node ℓ to neutral, and
  - one element corresponding to each line,
- as illustrated in Figure 6.10.
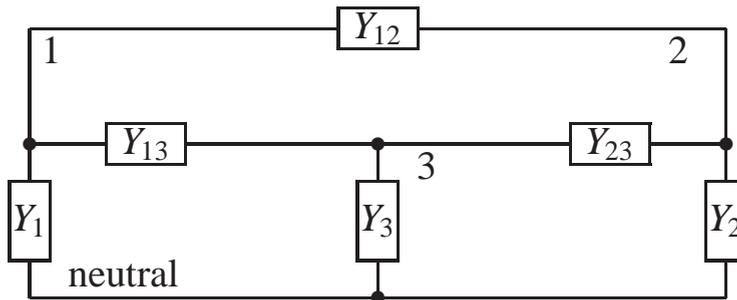
Fig. 6.10. Per-phase equivalent circuit model for three bus, three line system with parallel components combined.

Bus admittance matrix and power flow equations, continued

- Let us write Y_ℓ for the admittance of the element joining node ℓ to neutral, and
- $Y_{\ell k}$ for the admittance of the series element corresponding to a line joining buses ℓ and k.
- The series element is most easily characterized in terms of its impedance.
- For a series impedance $Z_{\ell k} = R_{\ell k} + X_{\ell k}\sqrt{-1}$ between buses ℓ and k, the corresponding admittance $Y_{\ell k}$ is given by:

$$
\begin{aligned}
Y_{\ell k} &= \frac{1}{Z_{\ell k}}, \\
&= \frac{1}{R_{\ell k} + X_{\ell k}\sqrt{-1}} \\
&= \frac{1}{R_{\ell k} + X_{\ell k}\sqrt{-1}} \times \frac{R_{\ell k} - X_{\ell k}\sqrt{-1}}{R_{\ell k} - X_{\ell k}\sqrt{-1}} \\
&= \frac{R_{\ell k} - X_{\ell k}\sqrt{-1}}{(R_{\ell k})^2 + (X_{\ell k})^2}.
\end{aligned}
\tag{6.7}
$$

Bus admittance matrix and power flow equations, continued

- Using Kirchhoff's laws, we can again obtain a relationship of the form $AV = I$ between current and voltage, where:

$$\forall \ell, k, A_{\ell k} = \begin{cases} Y_\ell + \sum_{k' \in \mathbb{J}(\ell)} Y_{\ell k'}, & \text{if } \ell = k, \\ -Y_{\ell k}, & \text{if } k \in \mathbb{J}(\ell) \text{ or } \ell \in \mathbb{J}(k), \\ 0, & \text{otherwise,} \end{cases} \tag{6.8}$$

- where $\mathbb{J}(\ell)$ is the set of buses joined directly by a transmission line to bus ℓ.
- A is called the **bus admittance matrix**.

6.2.2.5 *Generators and loads*

- When electricity is bought and sold, the power and energy are the quantities that are usually priced, not the voltage or current.
- However, real power does not completely describe the interaction between generators or loads and the system.
- We also have to characterize the injected **reactive power**.
- We can combine the real and reactive powers into the **complex power**, which is the sum of:

  the real power, and
  $\sqrt{-1}$ times the reactive power.

Generators and loads, continued

- The usefulness of this representation is that, for example, the complex power S_ℓ injected at node ℓ into the network is given by:

$$S_\ell = V_\ell I_\ell^*,$$

- where the superscript $*$ indicates **complex conjugate**.
- The current I_ℓ equals the sum of:

  the current flowing into the shunt element Y_ℓ, and
  the sum of the currents flowing into each line connecting ℓ to a bus
  $\quad k \in \mathbb{J}(\ell)$ through admittance $Y_{\ell k}$.

- We can substitute for the currents to obtain:

$$\begin{aligned}
S_\ell &= V_\ell[A_{\ell\ell}V_\ell + \sum_{k\in\mathbb{J}(\ell)} A_{\ell k}V_k]^*, \\
&= |V_\ell|^2 A_{\ell\ell}^* + \sum_{k\in\mathbb{J}(\ell)} A_{\ell k}^* V_\ell V_k^*.
\end{aligned} \tag{6.9}$$

Generators and loads, continued

- Let $A_{\ell k} = G_{\ell k} + B_{\ell k}\sqrt{-1}, \forall \ell, k$, where we note that by (6.7) and (6.8):
  - we have that $G_{\ell k} < 0$ and $B_{\ell k} > 0$ for $\ell \neq k$, and
  - we have that $G_{\ell \ell} > 0$ and the sign of $B_{\ell \ell}$ is indeterminate but typically less than zero;
- let $S_\ell = P_\ell + Q_\ell \sqrt{-1}, \forall \ell$, with:
  - for generator buses, $P_\ell > 0$ and Q_ℓ is typically positive,
  - for load buses, $P_\ell < 0$ and $Q_\ell < 0$;
- and let $V_\ell = u_\ell \exp(\theta_\ell \sqrt{-1}), \forall \ell$, with:
  - the voltage magnitude $u_\ell \approx 1$ in scaled units to satisfy voltage limits,
  - the voltage angle θ_ℓ typically between $-\pi/4$ and $\pi/4$ radians.

Generators and loads, continued

- Then we can separate (6.9) into real and imaginary parts:

$$P_\ell = \sum_{k \in \mathbb{J}(\ell) \cup \{\ell\}} u_\ell u_k [G_{\ell k} \cos(\theta_\ell - \theta_k) + B_{\ell k} \sin(\theta_\ell - \theta_k)], \quad (6.10)$$

$$Q_\ell = \sum_{k \in \mathbb{J}(\ell) \cup \{\ell\}} u_\ell u_k [G_{\ell k} \sin(\theta_\ell - \theta_k) - B_{\ell k} \cos(\theta_\ell - \theta_k)]. \quad (6.11)$$

- The equations (6.10) and (6.11), which are called the **power flow equality constraints**, must be satisfied at each bus ℓ.

Power balance

- A bus with a specified real and reactive power is called a *PQ* bus.
- We specify:
  - the real and reactive generations at the generator *PQ* buses according to the generator control settings, and
  - the real and reactive power at the load *PQ* buses according to supplied data.
- However, we cannot specify the injected power at all the buses since this would typically violate the first law of thermodynamics!

Reference bus

- A traditional, but *ad hoc* approach to finding a solution to the equations is to single out the reference bus.
- At this bus, instead of specifying injected real and reactive power, we specify the voltage magnitude.
- The reference generator is then assumed to produce whatever is needed to "balance" the real and reactive power for the rest of the system, assuming that such a solution exists.
- We re-interpret P_1 and Q_1 to be variables in our formulation and have eliminated these variables by writing them as a function of the rest of the variables.
- The reference bus supplies whatever power is necessary for power balance.

6.2.2.7 Non-linear equations

- We have n_{PQ} *PQ* buses, including both the *PQ* generators and the loads.
- Let $n = 2n_{PQ}$ and define a vector $x \in \mathbb{R}^n$ consisting of the voltage magnitudes and angles at the *PQ* buses.
- For every bus ℓ (that is, including the reference bus as well as the *PQ* buses) define functions $p_\ell : \mathbb{R}^n \to \mathbb{R}$ and $q_\ell : \mathbb{R}^n \to \mathbb{R}$ by:

$$\forall x \in \mathbb{R}^n, p_\ell(x) = \sum_{k \in \mathbb{J}(\ell) \cup \{\ell\}} u_\ell u_k [G_{\ell k} \cos(\theta_\ell - \theta_k) + B_{\ell k} \sin(\theta_\ell - \theta_k)] - P_\ell,$$

$$(6.12)$$

$$\forall x \in \mathbb{R}^n, q_\ell(x) = \sum_{k \in \mathbb{J}(\ell) \cup \{\ell\}} u_\ell u_k [G_{\ell k} \sin(\theta_\ell - \theta_k) - B_{\ell k} \cos(\theta_\ell - \theta_k)] - Q_\ell.$$

$$(6.13)$$

- The functions p_ℓ and q_ℓ represent the net real and reactive power flow, respectively, from bus ℓ into the rest of the system.
- Kirchhoff's laws require that the net real and reactive flow out of a bus must be zero.

Non-linear equations, continued

- Finally, define a vector function $g : \mathbb{R}^n \to \mathbb{R}^n$ that includes (6.12) and (6.13) for all the *PQ* buses, but omits (6.12) and (6.13) for the reference bus.

- We solve:

$$g(x) = \mathbf{0}. \tag{6.14}$$

- In summary, solving Kirchhoff's equations for the electric power network has been transformed into an equivalent problem:

  (i) solve (6.14), which is a system of non-linear simultaneous equations, and

  (ii) substitute into (6.10) and (6.11) for the reference bus.

6.2.3 Circuit changes

- If a real power injection changes at a bus ℓ then the entries in g corresponding to p_ℓ will change.
- If a reactive power injection changes at a bus ℓ then the entries in g corresponding to q_ℓ will change.
- If a transmission line between buses ℓ and k changes, then the entries of g corresponding to p_ℓ, q_ℓ, p_k, and q_k will change.
- The entries in the admittance matrix A will change in a manner analogous to the changes discussed in Section 4.1.3 for the DC circuit.

6.2.4 Problem characteristics

6.2.4.1 Number of variables and equations

- There are the same number of variables as equations in (6.14).

6.2.4.2 Non-existence of direct algorithms

- As with the non-linear circuit in Section 6.1, because the equations are non-linear, there is no direct algorithm to solve for x for arbitrary systems.

6.2.4.3 Number of solutions

- There may be no solutions, one solution, or even multiple solutions to (6.14).
- However, power systems are usually designed and operated so that the voltage magnitudes are near to nominal and the voltage angles are relatively close to $0°$.
- If we restrict our attention to solutions such that voltage magnitudes are all close to 1 (and make some other assumptions) then we can find conditions for the there to be at most one solution.

Symmetry

- The admittance matrix is symmetric.

Sparsity

- The matrix A is only sparsely populated with non-zero entries and each component of g depends on only a few components of x.

Values

- A typical line impedance has positive real and imaginary parts.
- The corresponding line admittance $Y_{\ell k}$ therefore has positive real part and negative imaginary part.
- If there is a line between buses ℓ and k then the entries $A_{\ell k} = G_{\ell k} + \sqrt{-1} B_{\ell k}$ in the admittance matrix satisfy $G_{\ell k} < 0, B_{\ell k} > 0$.
- The diagonal entries $A_{\ell \ell} = G_{\ell \ell} + \sqrt{-1} B_{\ell \ell}$ in the admittance satisfy $G_{\ell \ell} > 0$ and, typically, $B_{\ell \ell} < 0$.
- The resistance $R_{\ell k}$ of transmission lines is relatively small compared to the inductive reactance $X_{\ell k}$.
- Furthermore, the shunt elements are often also negligible compared to the inductive reactance.
- This means that:

$$\forall \ell, \forall k \in \mathbb{J}(\ell) \cup \{\ell\}, |G_{\ell k}| \ll |B_{\ell k}|.$$

7

Algorithms for non-linear simultaneous equations

Key issues

- Approximating non-linear functions by a **linear approximation**,
- using the linear approximation to improve our estimate of the solution,
- **convergence** of the sequence of iterates produced by repeated re-linearization,
- **variations** that reduce computational effort, and
- **sensitivity** and **large change analysis.**

7.1 Newton–Raphson method

- Consider a function $g : \mathbb{R}^n \to \mathbb{R}^n$ and suppose that we want to solve the simultaneous non-linear equations:

$$g(x) = \mathbf{0}. \tag{7.1}$$

7.1.1 Initial guess

- Let $x^{(0)}$ be the initial guess of a solution to (7.1).
- We seek an updated value of the vector $x^{(1)} = x^{(0)} + \Delta x^{(0)}$ such that:

$$g(x^{(1)}) = g(x^{(0)} + \Delta x^{(0)}) = \mathbf{0}. \tag{7.2}$$

7.1.2 Taylor approximation
7.1.2.1 Scalar function

$$
\begin{aligned}
g_1(x^{(1)}) &= g_1(x^{(0)} + \Delta x^{(0)}), \text{ since } x^{(1)} = x^{(0)} + \Delta x^{(0)}, \\
&\approx g_1(x^{(0)}) + \frac{\partial g_1}{\partial x_1}(x^{(0)})\Delta x_1^{(0)} + \cdots + \frac{\partial g_1}{\partial x_n}(x^{(0)})\Delta x_n^{(0)}, \\
&= g_1(x^{(0)}) + \sum_{k=1}^{n} \frac{\partial g_1}{\partial x_k}(x^{(0)})\Delta x_k^{(0)}, \\
&= g_1(x^{(0)}) + \frac{\partial g_1}{\partial x}(x^{(0)})\Delta x^{(0)}.
\end{aligned} \tag{7.3}
$$

- In (7.3), the symbol "$\approx$" should be interpreted to mean that the difference between the expressions to the left and to the right of the $\approx$ is small compared to $\left\| \Delta x^{(0)} \right\|$.

- Define the **remainder at the point** $x^{(0)}, e : \mathbb{R}^n \to \mathbb{R}$, by:

$$\forall \Delta x \in \mathbb{R}^n, e(\Delta x)$$
$$= g_1(x^{(0)} + \Delta x) - g_1(x^{(0)}) - \frac{\partial g_1}{\partial x_1}(x^{(0)})\Delta x_1 - \cdots - \frac{\partial g_1}{\partial x_n}(x^{(0)})\Delta x_n.$$

- By **Taylor's theorem with remainder**, if g_1 is partially differentiable with continuous partial derivatives then:

$$\lim_{\|\Delta x\| \to 0} \frac{e(\Delta x)}{\|\Delta x\|} = 0.$$

- As first mentioned in Section 2.6.3.5, the expression to the right of the $\approx$ in (7.3) is called a **first-order Taylor approximation**.
- For a partially differentiable function g_1 with continuous partial derivatives, the first-order Taylor approximation about $x = x^{(0)}$ approximates the behavior of g_1 in the vicinity of $x = x^{(0)}$.
- The first-order Taylor approximation represents a plane that is **tangential** to the graph of the function at the point $x^{(0)}$.

Scalar function, continued

- For example, suppose that $g_1 : \mathbb{R}^2 \to \mathbb{R}$ is defined by:

$$\forall x \in \mathbb{R}^2, g_1(x) = (x_1)^2 + (x_2)^2 + 2x_2 - 3.$$
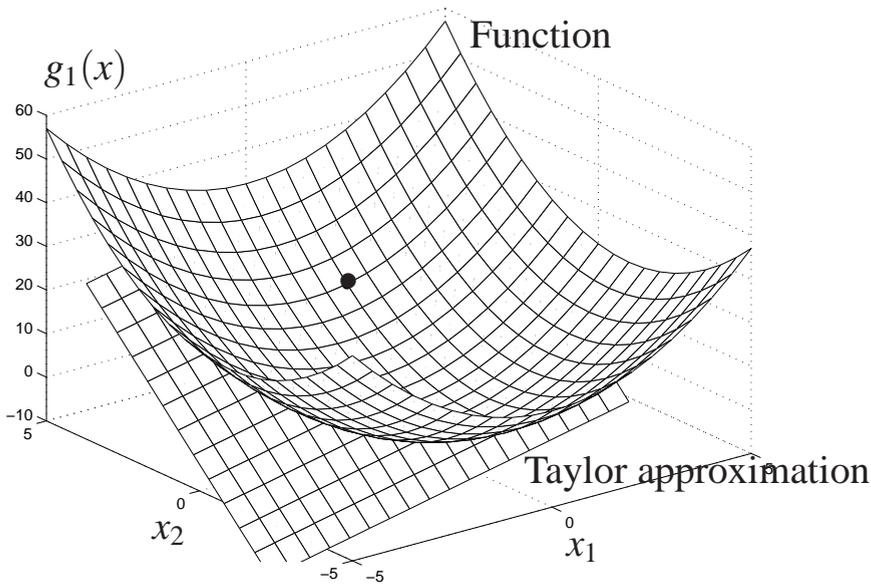
Fig. 7.1. Graph of function repeated from Figure 2.5 and its Taylor approximation about $x^{(0)} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$.

7.1.2.2 Vector function

- We now consider the vector function $g : \mathbb{R}^n \to \mathbb{R}^n$.
- Since g is a vector function and x is a vector, the Taylor approximation of g involves the $n \times n$ matrix of partial derivatives $\dfrac{\partial g}{\partial x}$ evaluated at $x^{(0)}$.

- A first-order Taylor approximation of g about $x^{(0)}$ yields:

$$g(x^{(0)} + \Delta x^{(0)}) \approx g(x^{(0)}) + \frac{\partial g}{\partial x}(x^{(0)})\Delta x^{(0)},$$

- where by the $\approx$ we mean that the norm of the difference between the expressions to the left and the right of $\approx$ is small compared to $\left\| \Delta x^{(0)} \right\|$.

Vector function, continued

- Define the remainder at the point $x^{(0)}$, $e : \mathbb{R}^n \to \mathbb{R}^n$, by:

$$\forall \Delta x \in \mathbb{R}^n, e(\Delta x)$$
$$= g(x^{(0)} + \Delta x) - g(x^{(0)}) - \frac{\partial g}{\partial x_1}(x^{(0)})\Delta x_1 - \cdots - \frac{\partial g}{\partial x_n}(x^{(0)})\Delta x_n.$$

- By **Taylor's theorem with remainder**, if g is partially differentiable with continuous partial derivatives then:

$$\lim_{\|\Delta x\| \to 0} \frac{\|e(\Delta x)\|}{\|\Delta x\|} = 0.$$

- The first-order Taylor approximation again represents a "plane" that is tangential to the graph of the function; however, the situation is much more difficult to visualize for a vector function.

7.1.2.3 Jacobian

- Recall from Section 2.5.3.2 that the matrix of partial derivatives is called the **Jacobian** and we will denote it by $J(\bullet)$.
- Using this notation, we have:

$$
\begin{aligned}
g(x^{(1)}) &= g(x^{(0)} + \Delta x^{(0)}), \text{by definition of } \Delta x^{(0)}, \\
&\approx g(x^{(0)}) + J(x^{(0)}) \Delta x^{(0)}.
\end{aligned}
\tag{7.4}
$$

- In some of our development, we will approximate the Jacobian when we evaluate the right-hand side of (7.4)
- In this case, the linear approximating function is no longer tangential to f.

7.1.3 Initial update

- Setting the right-hand side of (7.4) to zero to solve for $\Delta x^{(0)}$ yields a set of linear simultaneous equations:

$$J(x^{(0)})\Delta x^{(0)} = -g(x^{(0)}). \tag{7.5}$$

7.1.4 General update

$$
\begin{aligned}
J(x^{(\nu)})\Delta x^{(\nu)} &= -g(x^{(\nu)}), & (7.6) \\
x^{(\nu+1)} &= x^{(\nu)} + \Delta x^{(\nu)}. & (7.7)
\end{aligned}
$$

- (7.6)–(7.7) are called the **Newton–Raphson update**.
- $\Delta x^{(\nu)}$ is the **Newton–Raphson step direction**.

7.1.5 Discussion

- Three drawbacks of the Newton–Raphson method:

(i) The need to calculate the matrix of partial derivatives and solve a system of linear simultaneous equations at each iteration. Even with sparse matrix techniques, this can require considerable effort.

(ii) At some iteration we may find that the linear equation (7.6) does not have a solution, so that the update is not well-defined.

(iii) Even if (7.6) does have a solution at every iteration, the sequence of iterates generated may not converge to the solution of (7.1).

7.2 Variations on the Newton–Raphson method

- We will discuss various ways to reduce the effort involved in the basic Newton–Raphson method.

7.2.1 Approximation of the Jacobian

- Replace $J(x^{(\nu)})$ by a matrix $\tilde{J}^{(\nu)}$ such that, compared to using $J(x^{(\nu)})$ directly:

  (i) *LU* factorization of $\tilde{J}^{(\nu)}$ requires less effort (or has already been performed),

  (ii) an inverse of $\tilde{J}^{(\nu)}$ is easier to calculate, or

  (iii) evaluation of $\tilde{J}^{(\nu)}$ is more convenient.

- If the resulting approximation to the Newton–Raphson update satisfies suitable conditions, then it turns out that we will still iterate towards the solution.

7.2.1.1 The chord method

$$J(x^{(0)})\Delta x^{(\nu)} = -g(x^{(\nu)}), \tag{7.8}$$
$$x^{(\nu+1)} = x^{(\nu)} + \Delta x^{(\nu)}. \tag{7.9}$$

7.2.1.2 The Shamanskii method

$$J(x^{(N\lfloor \frac{\nu}{N} \rfloor)})\Delta x^{(\nu)} = -g(x^{(\nu)}), \tag{7.10}$$
$$x^{(\nu+1)} = x^{(\nu)} + \Delta x^{(\nu)}. \tag{7.11}$$

7.2.1.3 Approximating particular terms

- Replace small terms in the Jacobian by zero.

7.2.1.4 Analytic approximation to Jacobian

- We may have an *approximate* analytical model.
- Then we can combine a numerical evaluation of g with an approximate analytical model of J to use in the Newton–Raphson update.

7.2.1.5 Finite difference approximation to Jacobian

- The **forward difference approximation** between the point $x^{(v)}$ and the point $x^{(v)} + \Delta x$:

$$J(x^{(v)})\Delta x \approx g(x^{(v)} + \Delta x) - g(x^{(v)});$$

- The **central difference approximation** between the point $x^{(v)} - \Delta x$ and the point $x^{(v)} + \Delta x$:

$$2J(x^{(v)})\Delta x \approx g(x^{(v)} + \Delta x) - g(x^{(v)} - \Delta x);$$

or

- The **secant approximation**, for $x \in \mathbb{R}$, between the point $x^{(v)}$ and the point $x^{(v-1)}$:

$$\frac{\partial g}{\partial x}(x^{(v)}) \approx \frac{g(x^{(v)}) - g(x^{(v-1)})}{x^{(v)} - x^{(v-1)}}.$$

$g(x)$

$g(x^{(\nu)})$

x

$x^{(\nu-1)} \quad x^{(\nu)} - \Delta x \qquad x^{(\nu)} \qquad\qquad x^{(\nu)} + \Delta x$
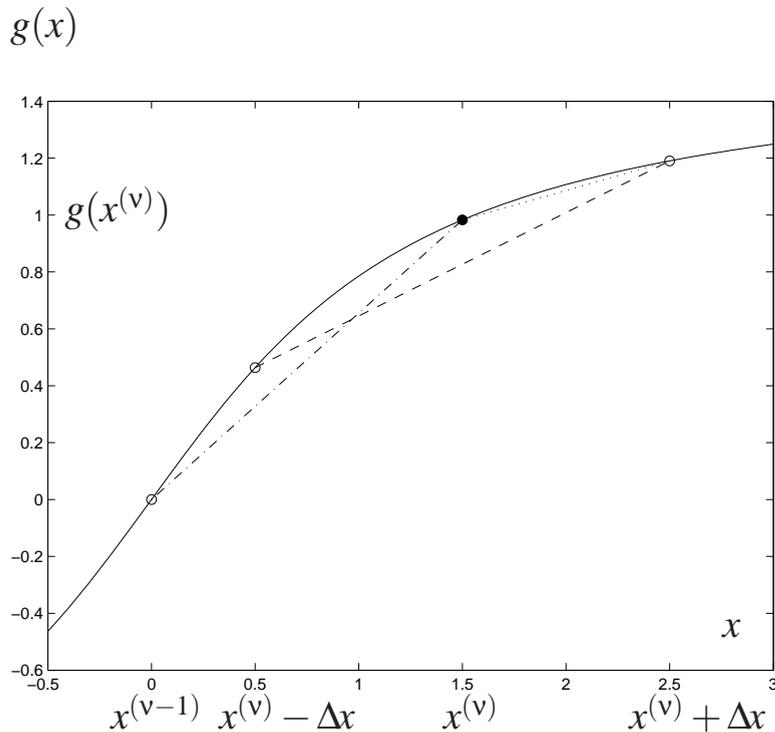
Fig. 7.2. Finite difference approximations to the derivative of a function $g : \mathbb{R} \to \mathbb{R}$ at a point $x^{(\nu)}$. The function g is illustrated as a solid curve. The point $\begin{bmatrix} x^{(\nu)} \\ g(x^{(\nu)}) \end{bmatrix} = \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}$ is indicated by the $\bullet$. The forward difference approximation with $\Delta x = 1$ is given by the slope of the dotted line. The central difference approximation with $\Delta x = 1$ is given by the slope of the dashed line. The secant approximation for $x^{(\nu-1)} = 0$ is given by the slope of the dash-dotted line.

7.2.1.6 *Quasi-Newton methods*

- Consider a first-order Taylor approximation of g about $x^{(\nu-1)}$:

$$g(x^{(\nu-1)} + \Delta x^{(\nu-1)}) \approx g(x^{(\nu-1)}) + J(x^{(\nu-1)})\Delta x^{(\nu-1)}.$$

- Substituting from the Newton–Raphson update equations (7.6)–(7.7) applied to calculate $x^{(\nu)}$, we obtain:

$$g(x^{(\nu)}) \approx g(x^{(\nu-1)}) + J(x^{(\nu-1)})(x^{(\nu)} - x^{(\nu-1)}).$$

- Re-arranging, we have:

$$J(x^{(\nu-1)})(x^{(\nu)} - x^{(\nu-1)}) \approx g(x^{(\nu)}) - g(x^{(\nu-1)}). \qquad (7.12)$$

- **Quasi-Newton methods** involve successively updating each approximation $\tilde{J}^{(\nu-1)}$ so that the updated approximation $\tilde{J}^{(\nu)}$ used for calculating $x^{(\nu+1)}$ satisfies the **quasi-Newton condition:**

$$\forall \nu > 0, \tilde{J}^{(\nu)}(x^{(\nu)} - x^{(\nu-1)}) = g(x^{(\nu)}) - g(x^{(\nu-1)}). \qquad (7.13)$$

- Quasi-Newton methods generalize the secant approximation to functions $g : \mathbb{R}^n \to \mathbb{R}^n$.
- The approximation $\tilde{J}^{(\nu)}$, (which is used in the calculation of $x^{(\nu+1)}$) is chosen to mimic the behavior of the change in g that resulted from the choice of $x^{(\nu)}$ in the previous iteration.

- Under mild conditions, if $\tilde{J}^{(\nu-1)}$ is symmetric then **symmetric rank two updates** can be found that satisfy the Quasi-Newton condition.

7.2.2 Iterative algorithms

- If the Jacobian is large and non-sparse, then the factorization- or inversion-based techniques that we have discussed so far may not be effective.
- Iterative algorithms may be used.

7.2.3 Pre-conditioning

- Pre-conditioning can be used to help with the solution of the update equation if an approximate inverse to the Jacobian is known.
- A simple "pre-conditioner" is the diagonal matrix consisting of the inverse of the diagonal elements of the Jacobian.
- Pre-conditioning is often used in combination with iterative methods.

7.2.4 Automatic differentiation

- If the calculation of g is performed by code that implements a direct algorithm, it is possible to systematically transform the code for calculating g into code that calculates the Jacobian.

7.3 Local convergence of iterative methods

7.3.1 Closeness to a solution

- In this section, we discuss three measures of closeness to a solution that are candidates for use as a stopping criterion.
- We then discuss using the iteration count and the combination of several stopping criteria.

7.3.1.1 Function value

$$\left\| g(x^{(\nu)}) \right\|_{\infty} \leq \varepsilon_g, \tag{7.14}$$

$$\left\| g(x^{(\nu)}) \right\|_{\infty} \leq \varepsilon_g \left\| g(x^{(0)}) \right\|_{\infty}. \tag{7.15}$$

7.3.1.2 Iteration space

$$\left\| x^{(\nu)} - x^{\star} \right\|_2 \;\leq\; \varepsilon_x, \tag{7.16}$$

$$\left\| x^{(\nu)} - x^{\star} \right\|_2 \;\leq\; \varepsilon_x \left\| x^{(0)} - x^{\star} \right\|_2 .$$

7.3.1.3 Change in iterate

$$\left\| \Delta x^{(\nu)} \right\| \leq \varepsilon_{\Delta x}. \tag{7.17}$$

7.3.1.4 Iteration count

- It is common to limit the total number of iterations.

7.3.1.5 Combined stopping criteria

- Combinations of criteria are used in practice to balance the desire to:
  - get close to a solution, but
  - not perform an excessive number of iterations.

7.3.2 *The Cauchy criterion and contraction mappings*
7.3.2.1 *Cauchy sequences*

Definition 7.1 A sequence $\{x^{(\nu)}\}_{\nu=0}^{\infty}$ is said to be a **Cauchy sequence** or **Cauchy** if:

$$\forall \varepsilon \in \mathbb{R}_{++}, \exists N \in \mathbb{Z}_+ \text{ such that } (\nu, \nu' \in \mathbb{Z}_+ \text{ and } \nu, \nu' \geq N) \Rightarrow \left(\left\| x^{(\nu)} - x^{(\nu')} \right\| \leq \varepsilon \right).$$

$\square$

- The weaker condition:

$$\forall \varepsilon \in \mathbb{R}_{++}, \exists N \in \mathbb{Z}_+ \text{ such that } (\nu \in \mathbb{Z}_+ \text{ and } \nu \geq N) \Rightarrow \left(\left\| x^{(\nu+1)} - x^{(\nu)} \right\| \leq \varepsilon \right),$$
$$(7.18)$$

- is insufficient to guarantee that the sequence $\{x^{(\nu)}\}_{\nu=0}^{\infty}$ is Cauchy.

Lemma 7.1 *A sequence $\{x^{(\nu)}\}_{\nu=0}^{\infty}$ of real vectors converges to a limit in $\mathbb{R}^n$ if and only if it is Cauchy.* $\square$

Definition 7.2 A function $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ (or $\Phi : \mathbb{R}^n \to \mathbb{R}^{m \times n}$) is **Lipschitz continuous**:

- on a set $\mathbb{S} \subseteq \mathbb{R}^n$,
- with respect to a norm $\|\bullet\|$ on the domain $\mathbb{R}^n$,
- with respect to a norm $\|\bullet\|$ on the range $\mathbb{R}^m$ (or to a norm on $\mathbb{R}^{m \times n}$), and
- with constant $L \geq 0$, if:

$$\forall x, x' \in \mathbb{S}, \left\|\Phi(x) - \Phi(x')\right\| \leq L \left\|x - x'\right\|. \tag{7.19}$$
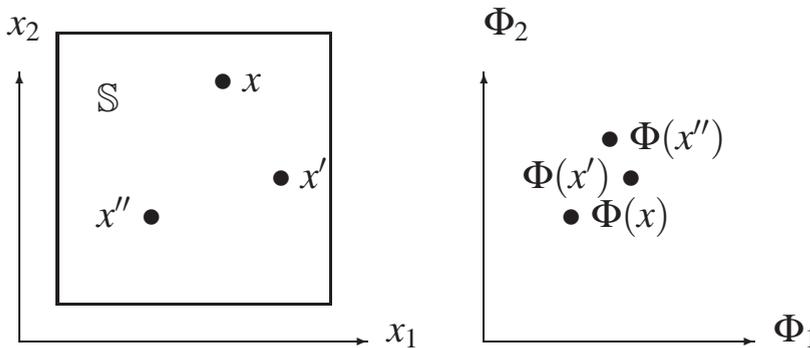
$\square$

Fig. 7.3. Points x, x', and x'' in a set $\mathbb{S} \subseteq \mathbb{R}^2$ (left panel) and their images $\Phi(x), \Phi(x')$, and $\Phi(x'')$ (right panel) under a Lipschitz continuous function $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$.

7.3.2.3 Contraction mapping

Definition 7.3 A map $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ is called a **contraction mapping** or a **contraction map**:

- on a set $\mathbb{S} \subseteq \mathbb{R}^n$, and
- with respect to a norm $\|\bullet\|$ on $\mathbb{R}^n$,

if $\exists \, 0 \leq L < 1$ such that:

$$\forall x, x' \in \mathbb{S}, \left\| \Phi(x) - \Phi(x') \right\| \leq L \left\| x - x' \right\|.$$

$\square$

- A map from $\mathbb{R}^n$ to $\mathbb{R}^n$ is a contraction map on $\mathbb{S} \subseteq \mathbb{R}^n$ if it is:
  - Lipschitz continuous on $\mathbb{S}$ for one particular norm applied to both its domain and range, and
  - the Lipschitz constant is less than one.
- The map Φ illustrated in Figure 7.3 is a contraction mapping with respect to the Euclidean norm.

7.3.2.4 *General iterative methods and fixed points*

- Consider a general iterative method:

$$\forall \nu \in \mathbb{Z}_+, x^{(\nu+1)} = \Phi(x^{(\nu)}), \qquad (7.20)$$

- where $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ represents the calculations during a single iteration.

Definition 7.4 A point $x^\star$ is called a **fixed point** of a map $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ if $x^\star = \Phi(x^\star)$. $\square$
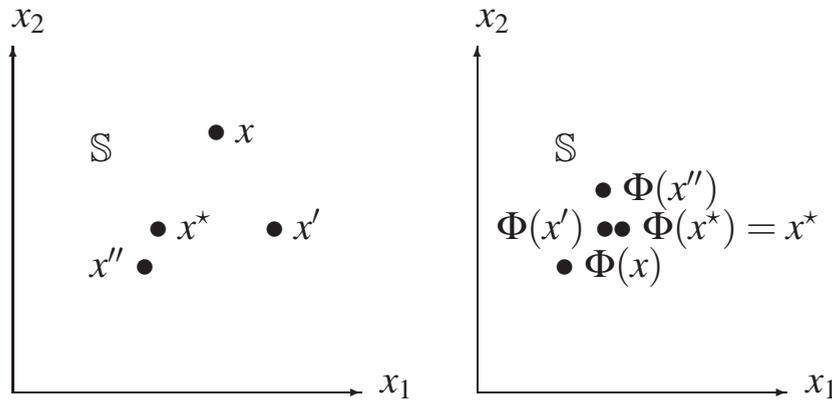

Fig. 7.4. Points x, x', x'', and $x^\star$ in $\mathbb{R}^2$ (left panel) and their images $\Phi(x), \Phi(x'), \Phi(x'')$, and $\Phi(x^\star)$ (right panel) under a function $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$. The point $x^\star$ is a fixed point of Φ because $\Phi(x^\star) = x^\star$.

7.3.2.5 Contraction mapping theorem

Theorem 7.2 *Suppose that $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ is a contraction mapping with Lipschitz constant $0 \le L < 1$ with respect to some norm $\|\bullet\|$ on a closed set $\mathbb{S} \subseteq \mathbb{R}^n$. Also suppose that $\forall x \in \mathbb{S}, \Phi(x) \in \mathbb{S}$. Then, there exists a unique $x^\star \in \mathbb{S}$ that is a fixed point of Φ. Moreover, for any $x^{(0)} \in \mathbb{S}$, the sequence of iterates generated by the iterative method (7.20) converges to $x^\star$ and satisfies the bound:*

$$\forall \nu \in \mathbb{Z}_+, \left\| x^{(\nu)} - x^\star \right\| \le (L)^\nu \left\| x^{(0)} - x^\star \right\|. \tag{7.21}$$

Proof The long proof is divided into four parts:

  (i) proving that $\{x^{(\nu)}\}_{\nu=0}^{\infty}$ is Cauchy and has a limit that is contained in $\mathbb{S}$;

 (ii) proving that the limit is a fixed point of Φ;

(iii) proving that the fixed point is unique; and

(iv) proving that the sequence converges to the fixed point according to (7.21).

$\square$

7.3.3 The chord and Newton–Raphson methods
7.3.3.1 The chord method

Theorem 7.3 *Consider a function $g : \mathbb{R}^n \to \mathbb{R}^n$. Let $\|\bullet\|$ be a norm on $\mathbb{R}^n$ and let $\|\bullet\|$ also stand for the corresponding induced matrix norm. Suppose that there exist $a, b, c,$ and $\overline{\rho} \in \mathbb{R}_+$ such that:*

(i) *g is partially differentiable with continuous partial derivatives at $x^{(0)}$, with Jacobian $J(x^{(0)})$ satisfying:*

$$\left\| [J(x^{(0)})]^{-1} \right\| \leq a,$$
$$\left\| [J(x^{(0)})]^{-1} g(x^{(0)}) \right\| \leq b,$$

(ii) *g is partially differentiable in a closed ball of radius $\overline{\rho}$ about $x^{(0)}$, with Jacobian J that is Lipschitz continuous with Lipschitz constant c. That is,*

$$\forall x, x' \in \left\{ x \in \mathbb{R}^n \,\middle|\, \left\| x - x^{(0)} \right\| \leq \overline{\rho} \right\}, \, \left\| J(x) - J(x') \right\| \leq c \left\| x - x' \right\|.$$

(iii) *$abc < \frac{1}{2}$, and*

(iv) $\rho_- \le \overline{\rho}$ *where* $\rho_- = \frac{1-\sqrt{1-2abc}}{ac}$.

Then:

(i) *In the open ball of radius* $\rho_+ = \min\left\{\overline{\rho}, \left(1+\sqrt{1-2abc}\right)/(ac)\right\}$ *about* $x^{(0)}$ *there is a unique solution* $x^\star$ *of* $g(x) = \mathbf{0}$. *(There may be other solutions outside this ball.)*

(ii) *Consider the chord update (7.8)–(7.9) with* $x^{(0)}$ *as initial guess. The sequence of iterates converges to* $x^\star$ *and each iterate* $x^{(\nu)}$ *is contained in the closed ball of radius* ρ_- *about* $x^{(0)}$. *Furthermore,*

$$\forall \nu \in \mathbb{Z}_+, \left\| x^{(\nu)} - x^\star \right\| \le (ac\rho_-)^\nu \rho_-. \qquad (7.22)$$

The chord method, continued

x_2

$\{x \in \mathbb{R}^2 | g_1(x) = 0\}$

$x^{(0)}$

$x^\star$
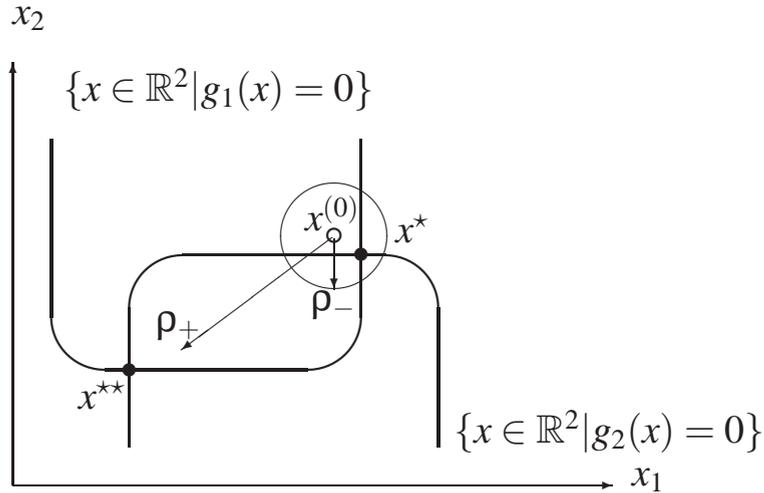
ρ_-

ρ_+

$x^{\star\star}$

$\{x \in \mathbb{R}^2 | g_2(x) = 0\}$

x_1
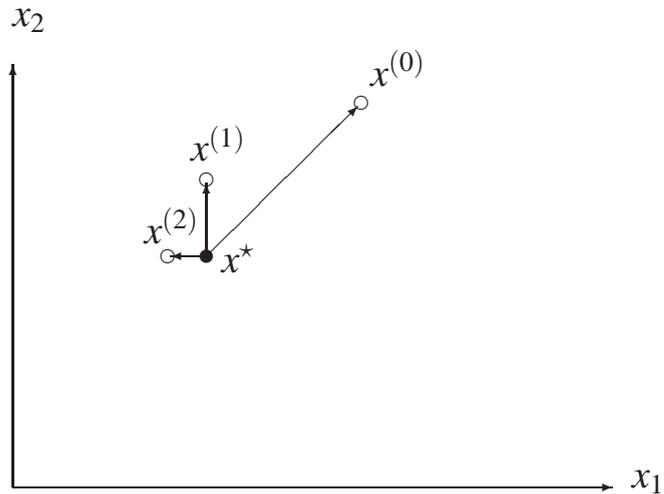
Fig. 7.5. Illustration of chord and Kantorovich theorems.

Fig. 7.6. Illustration of the linear rate of convergence in chord theorem.

Proof

- We define $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ to be the map that represents the update in the chord method.

$$\forall x \in \mathbb{R}^n, \Phi(x) = x - \left[J(x^{(0)})\right]^{-1} g(x).$$

- The proof is divided into four parts:

  (i) proving that the iterates stay in $\mathbb{S} = \left\{ x \in \mathbb{R}^n \,\middle|\, \left\| x - x^{(0)} \right\| \leq \rho_- \right\}$,

  (ii) proving that Φ is a contraction map with Lipschitz constant $L = ac\rho_- < 1$ so that, by the contraction mapping Theorem 7.2, there exists a unique $x^\star \in \mathbb{S}$ that is a fixed point of Φ,

  (iii) proving that the fixed point $x^\star$ of Φ satisfies (7.1) and (7.22), and

  (iv) proving that $x^\star$ is the only solution within a distance ρ_+ of $x^{(0)}$.

$\square$

- The rate of convergence is linear.

7.3.3.2 Kantorovich theorem

Theorem 7.4 *(Kantorovich) Consider a function $g : \mathbb{R}^n \to \mathbb{R}^n$. Let $\|\bullet\|$ be a norm on $\mathbb{R}^n$ and let $\|\bullet\|$ also stand for the corresponding induced matrix norm. Suppose that there exists $a, b, c,$ and $\overline{\rho} \in \mathbb{R}_+$ such that:*

(i) *g is partially differentiable with continuous partial derivatives at $x^{(0)}$, with Jacobian $J(x^{(0)})$ satisfying:*

$$\left\| [J(x^{(0)})]^{-1} \right\| \leq a,$$
$$\left\| [J(x^{(0)})]^{-1} g(x^{(0)}) \right\| \leq b,$$

(ii) *g is partially differentiable, with Jacobian J that is Lipschitz continuous with Lipschitz constant c in a closed ball of radius $\overline{\rho}$ about $x^{(0)}$. That is,*

$$\forall x, x' \in \left\{ x \in \mathbb{R}^n \,\middle|\, \left\| x - x^{(0)} \right\| \leq \overline{\rho} \right\}, \left\| J(x) - J(x') \right\| \leq c \left\| x - x' \right\|.$$

(iii) *$abc < \frac{1}{2},$ and*

(iv) $\rho_- \leq \bar{\rho}$ *where* $\rho_- = \frac{1-\sqrt{1-2abc}}{ac}$.

Then:

(i) *In the open ball of radius* $\rho_+ = \min\left\{\bar{\rho}, \frac{1+\sqrt{1-2abc}}{ac}\right\}$ *about* $x^{(0)}$, *there is only one solution* $x^\star$ *of* $g(x) = 0$. *(There may be other solutions outside this ball.)*

(ii) *Consider the Newton–Raphson update (7.6)–(7.7) with* $x^{(0)}$ *as initial guess. The sequence of iterates converges to* $x^\star$ *and each iterate* $x^{(\nu)}$ *is contained in the closed ball of radius* ρ_- *about* $x^{(0)}$. *Furthermore,*

$$\forall \nu \in \mathbb{Z}_+, \left\| x^{(\nu)} - x^\star \right\| \leq \frac{(2abc)^{((2)^\nu)}}{(2)^\nu ac}. \tag{7.23}$$

□

- The rate of convergence is quadratic.

7.3.3.3 Discussion

- The chord theorem and the Kantorovich theorem are "local."
- If the Jacobian is non-singular at the initial guess (so that a is well-defined),
- if the initial guess satisfies the equations sufficiently well (so that the norm b of the initial update:

$$\begin{aligned} b &= \left\| [J(x^{(0)})]^{-1} g(x^{(0)}) \right\|, \\ &= \left\| \Delta x^{(0)} \right\|, \end{aligned}$$

is small), and
- if the Jacobian does not vary too much over the closed ball of radius $\bar{\rho}$ about $x^{(0)}$ (so that c is small),
- then the chord and the Newton–Raphson updates converge to the solution.

7.3.4 Computational effort

- Suppose $\overline{\rho}$ is the best bound we have on the initial error; that is:

$$\left\| x^{(0)} - x^\star \right\| \leq \overline{\rho}. \tag{7.24}$$

- We want to estimate the number of iterations N such that the error bound is reduced by a factor $\varepsilon_x < 1$ so that:

$$\left\| x^{(N)} - x^\star \right\| \leq \varepsilon_x \overline{\rho}. \tag{7.25}$$

7.3.4.1 Chord method

- The computations required for N iterations are:
  - one evaluation and one factorization of the Jacobian, requiring effort on the order of $(n)^3$, and
  - one evaluation of g per iteration, one forwards and backwards substitution per iteration, and one vector addition per iteration, requiring effort on the order of $N(n)^2$.
- The overall effort is on the order of $(n)^3 + N(n)^2$ and the average effort per iteration is on the order of $(n)^3/N + (n)^2$.
- We must find a bound on the size of N that is necessary to satisfy (7.25).

$$
\begin{aligned}
\left\| x^{(N)} - x^\star \right\| &\leq (ac\rho_-)^N \rho_-, \\
&= (ac\rho_-)^N \left(\frac{\rho_-}{\overline{\rho}} \right) \overline{\rho}, \\
&\leq (ac\rho_-)^N \left(\frac{\rho_-}{\rho_+} \right) \overline{\rho},
\end{aligned}
$$

- since $\rho_+ \leq \overline{\rho}$ by definition.

Chord method, continued

- Then (7.25) will be satisfied if $(ac\rho_-)^N \left(\frac{\rho_-}{\rho_+}\right) \le \varepsilon_x$.

- Re-arranging this condition we obtain that $(ac\rho_-)^N \le \dfrac{\varepsilon_x \rho_+}{\rho_-}$.

- Taking natural logarithms and re-arranging, we obtain
$$N \ge \frac{\ln(\varepsilon_x) + \ln(\rho_+) - \ln(\rho_-)}{\ln(ac\rho_-)},$$

- noting that $\ln(ac\rho_-) < 0$.

- Overall effort is on the order of:
$$(n)^3 + \frac{\ln(\varepsilon_x) + \ln(\rho_+) - \ln(\rho_-)}{\ln(ac\rho_-)} (n)^2.$$

- Computational effort grows with $(n)^3$ and $(n)^2 |\ln(\varepsilon_x)|$.

7.3.4.2 Newton–Raphson method

- The computations required for N iterations are:
  - one evaluation and factorization of the Jacobian per iteration, requiring effort on the order of $N(n)^3$, and
  - one evaluation of g per iteration and one forwards and backwards substitution per iteration, requiring effort on the order of $N(n)^2$.
- The overall effort is on the order of $N(n)^3$.
- Again, we must find a bound on the size of N that is necessary to satisfy (7.25).

Newton–Raphson method, continued

$$\left\| x^{(N)} - x^\star \right\| \leq \frac{(2abc)^{((2)^N)}}{(2)^N ac},$$

$$= \frac{(2abc)^{((2)^N)}}{(2)^N ac\overline{\rho}}\overline{\rho},$$

$$\leq \frac{(2abc)^{((2)^N)}}{(2)^N ac\rho_+}\overline{\rho},$$

$$\leq \frac{(2abc)^{((2)^N)}}{ac\rho_+}\overline{\rho},$$

- since $\rho_+ \leq \overline{\rho}$ by definition and $(2)^N \geq 1$.

Newton–Raphson method, continued

- Then (7.25) will be satisfied if $\dfrac{(2abc)^{((2)^N)}}{ac\rho_+} \leq \varepsilon_x$.

- Re-arranging this condition we obtain that $(2abc)^{((2)^N)} \leq ac\rho_+\varepsilon_x$.

- Taking natural logarithms, we obtain $((2)^N)\ln(2abc) \leq \ln(ac\rho_+\varepsilon_x)$.

- Now $2abc < 1$ by hypothesis, so $\ln(2abc) < 0$ and dividing both sides by the negative number $\ln(2abc)$ yields $(2)^N \geq \dfrac{\ln(ac\rho_+\varepsilon_x)}{\ln(2abc)}$.

- Taking natural logarithms again and re-arranging yields:

$$N \geq \frac{\ln\left(\frac{\ln(ac\rho_+\varepsilon_x)}{\ln(2abc)}\right)}{\ln(2)} = \frac{\ln(|\ln(ac\rho_+\varepsilon_x)|) - \ln(|\ln(2abc)|)}{\ln(2)}.$$

- Overall effort is:

$$(n)^3\frac{\ln(|\ln(ac\rho_+\varepsilon_x)|) - \ln(|\ln(2abc)|)}{\ln(2)}.$$

- For small ε_x the computational effort grows with $(n)^3\ln(|\ln(\varepsilon_x)|)$.

7.3.4.3 Quasi-Newton methods

- Assuming super-linear convergence we again find that the number of iterations N grows with $\ln(|\ln(\varepsilon_x)|)$ and consequently the computational effort grows with $(n)^2 \ln(|\ln(\varepsilon_x)|)$.
- This effort grows much more slowly with n than for the Newton–Raphson method.

7.3.4.4 Other variations

- Often, the variations that avoid a complete factorization at every iteration will be more attractive than the basic Newton–Raphson method.

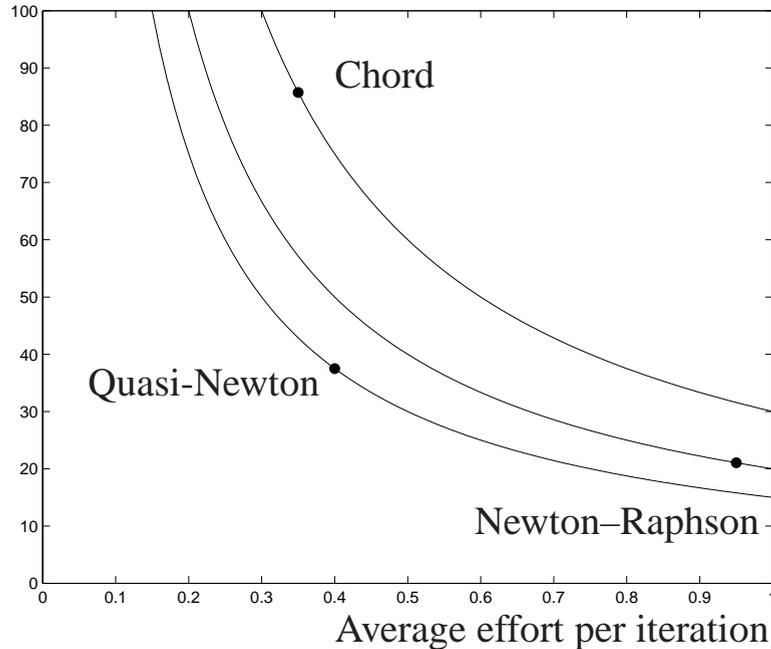Number of iterations to satisfy stopping criterion

Fig. 7.7. The qualitative tradeoff between effort per iteration and number of iterations.

7.3.4.6 Calculation of Jacobian

- The analysis so far has assumed that the entries of J take no more effort to calculate than the entries of g.
- It is sometimes more difficult to calculate entries of J than it is to calculate entries of g.
- In this case, we may choose to use a method that uses less information about J but also has a slower rate of convergence, and consequently a larger required value of N, because of the savings in the computational effort per iteration.

7.3.5 Discussion

- The chord method and the Newton–Raphson method have great local performance.
- The theorems can provide qualitative insights into convergence.
- However, we also consider cases where the initial guess is far from the solution.

7.4 Globalization procedures

- We must safeguard our algorithm from two related issues:

  (i) singular Jacobian, and
  (ii) excessively large steps.

7.4.1 Singular Jacobian

7.4.1.1 Example

$$\forall x \in \mathbb{R}, g(x) = (x-2)^3 + 1.$$
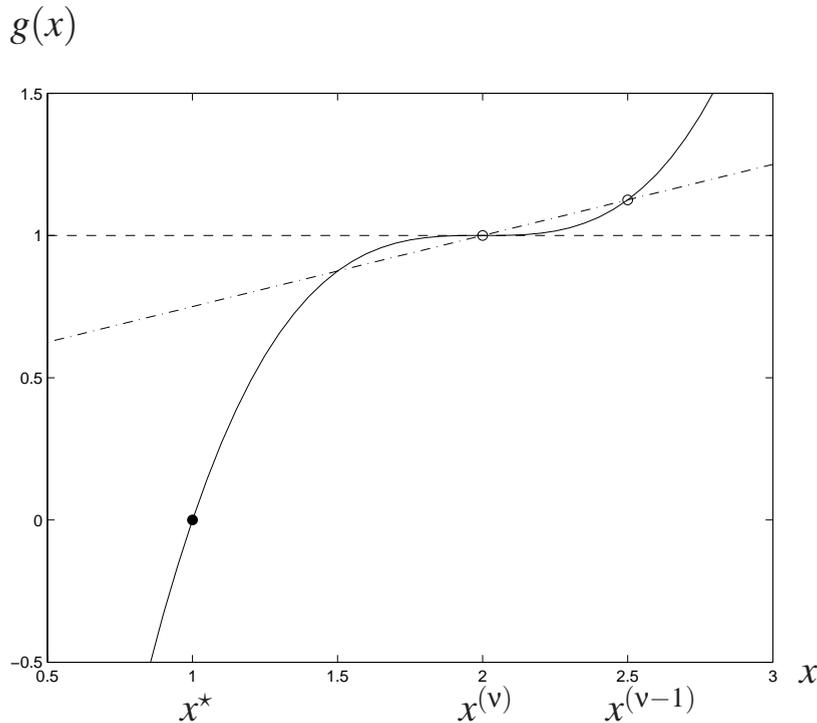
$g(x)$

Fig. 7.8. A function with a singular Jacobian at the point $x^{(\nu)} = 2$. The first-order Taylor approximation about $x^{(\nu)}$ is shown dashed. The approximation implied by the secant approximation through $x^{(\nu)}$ and $x^{(\nu-1)}$ is shown as the dot-dashed line.

7.4.1.2 Modified factorization

- If J is singular at any iterate then the basic Newton–Raphson update will fail.
- An *ad hoc* approach to this problem is to modify terms in $J(x)$ if it is singular and then solve the resulting update equation.
- For example, for $g : \mathbb{R} \to \mathbb{R}$, if $|J(x^{(\nu)})| < E$ for some threshold $E \in \mathbb{R}_{++}$, then we might replace $J(x^{(\nu)})$ by the secant approximation:

$$\tilde{J}^{(\nu)} = \frac{g(x^{(\nu)}) - g(x^{(\nu-1)})}{x^{(\nu)} - x^{(\nu-1)}}$$

- or replace $J(x^{(\nu)})$ by the value E.
- For $g : \mathbb{R}^n \to \mathbb{R}^n$, during factorization of J, if we encounter a small or zero pivot, we simply replace the pivot by a small non-zero number.

7.4.2 Step-size selection
7.4.2.1 Region of validity of approximation of function
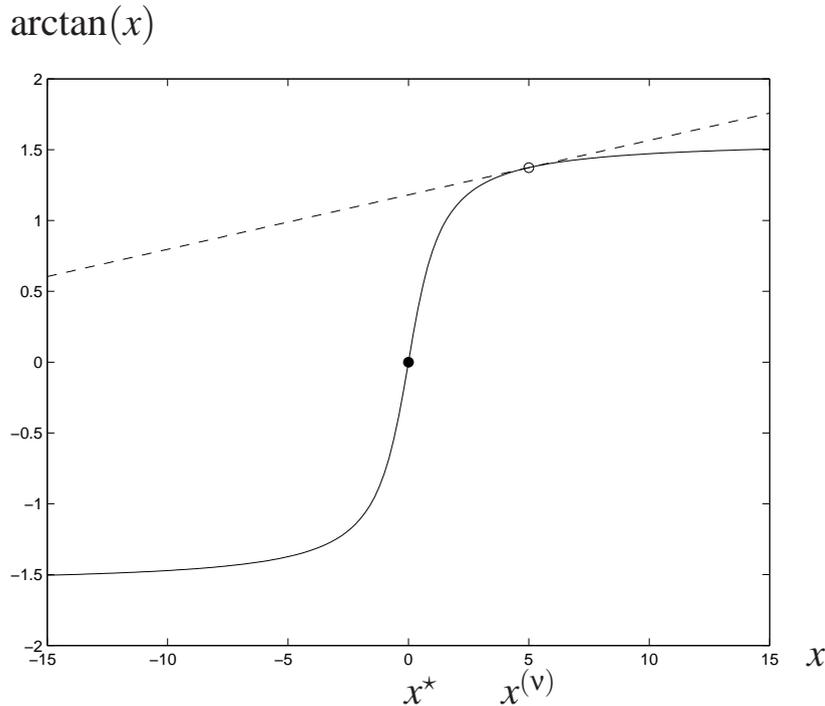
$\arctan(x)$

Fig. 7.9. The inverse tan function (shown solid) and its first-order Taylor approximation about $x^{(\nu)} = 5$ (shown dashed.) The point
$$\begin{bmatrix} x^{(\nu)} \\ g(x^{(\nu)}) \end{bmatrix} = \begin{bmatrix} 5 \\ 1.3734 \end{bmatrix}$$
is illustrated with a $\circ$, while the solution to the equation $g(x) = \mathbf{0}$ is shown with a $\bullet$.

Region of validity of approximation of function, continued

- Consider the iterate $x^{(\nu)}$ shown in Figure 7.9.
- For the function shown, $\left\|x^{(\nu)} - x^\star\right\| = \|5 - 0\| = 5$.
- Using step-size equal to 1:

$$\left\|x^{(\nu+1)} - x^\star\right\| > \left\|x^{(\nu)} - x^\star\right\|,$$
$$\left\|g(x^{(\nu+1)})\right\| > \left\|g(x^{(\nu)})\right\|.$$

- If the Newton–Raphson step direction $\Delta x^{(\nu)}$ is so large that it would take the next iterate outside the region of validity of the linear approximation, then we should not move as far as $\Delta x^{(\nu)}$ suggests.
- Instead, we should consider moving a smaller step in the direction of $\Delta x^{(\nu)}$.

7.4.2.2 Step-size rules

- **Damped Newton method**: pick a fixed $0 < \alpha < 1$ and modify (7.7) to:

$$x^{(\nu+1)} = x^{(\nu)} + \alpha \Delta x^{(\nu)}.$$

- Allow the step-size to vary with iteration:

$$x^{(\nu+1)} = x^{(\nu)} + \alpha^{(\nu)} \Delta x^{(\nu)}, \tag{7.26}$$

- where $0 < \alpha^{(\nu)} \leq 1$ is chosen at each iteration so that:

$$\left\| g(x^{(\nu)} + \alpha^{(\nu)} \Delta x^{(\nu)}) \right\| < \left\| g(x^{(\nu)}) \right\|. \tag{7.27}$$

- If the L_2 norm is chosen in (7.27) then it is possible to choose a suitable $\alpha^{(\nu)}$ if:
  - g is partially differentiable with continuous partial derivatives, and
  - the step direction $\Delta x^{(\nu)}$ satisfies:

$$[\Delta x^{(\nu)}]^\dagger J(x^{(\nu)})^\dagger g(x^{(\nu)}) < 0. \tag{7.28}$$

7.4.2.3 Armijo step-size rule

- Condition (7.27) does not specify by *how much* the norm of g should decrease to ensure that we obtain a satisfactory improvement in the satisfaction of the equations.
- A variation on (7.27) that does specify a "sufficient" decrease requires that:

$$\left\| g(x^{(\nu)} + \alpha^{(\nu)} \Delta x^{(\nu)}) \right\| \leq (1 - \delta \alpha^{(\nu)}) \left\| g(x^{(\nu)}) \right\|, \qquad (7.29)$$

- where $0 < \delta < 1$ is a positive constant.
- To understand (7.29), suppose that $\alpha^{(\nu)}$ is small enough so that the linear Taylor approximation is accurate and also assume that the Newton–Raphson step direction was used:

$$
\begin{aligned}
g(x^{(\nu)} &+ \alpha^{(\nu)} \Delta x^{(\nu)}) \\
&\approx \ g(x^{(\nu)}) + J(x^{(\nu)}) \alpha^{(\nu)} \Delta x^{(\nu)}, \text{ since } \alpha^{(\nu)} \text{ is assumed to be} \\
&\qquad \text{small enough so that the linear Taylor approximation is accurate,} \\
&= \ g(x^{(\nu)}) - \alpha^{(\nu)} g(x^{(\nu)}), \text{ by definition of } \Delta x^{(\nu)}, \\
&= \ (1 - \alpha^{(\nu)}) g(x^{(\nu)}).
\end{aligned}
$$

- Therefore, taking norms:

$$\left\| g(x^{(\nu)} + \alpha^{(\nu)}\Delta x^{(\nu)}) \right\| \approx (1 - \alpha^{(\nu)}) \left\| g(x^{(\nu)}) \right\|.$$

- With a step-size of $\alpha^{(\nu)}$, the best we could expect is for $\left\| g(x^{(\nu)} + \alpha^{(\nu)}\Delta x^{(\nu)}) \right\|$ to be reduced by a factor of $(1 - \alpha^{(\nu)})$ compared to $\left\| g(x^{(\nu)}) \right\|$.

- In practice, we will not achieve this reduction, as allowed for in condition (7.29).

- Condition (7.29) together with a reduction rule for choosing $\alpha^{(\nu)}$ is called the **Armijo step-size rule**.

- For example, the rule could be to find the largest step-size of the form:

$$\alpha^{(\nu)} = (2)^{-k}, k \geq 0, \tag{7.30}$$

- that satisfies (7.29).

x_2

$x^{(\nu)} + \Delta x^{(\nu)}$

$x^{(\nu)} + (2)^{-1} \times \Delta x^{(\nu)}$

$x^{(\nu)} + (2)^{-2} \times \Delta x^{(\nu)}$
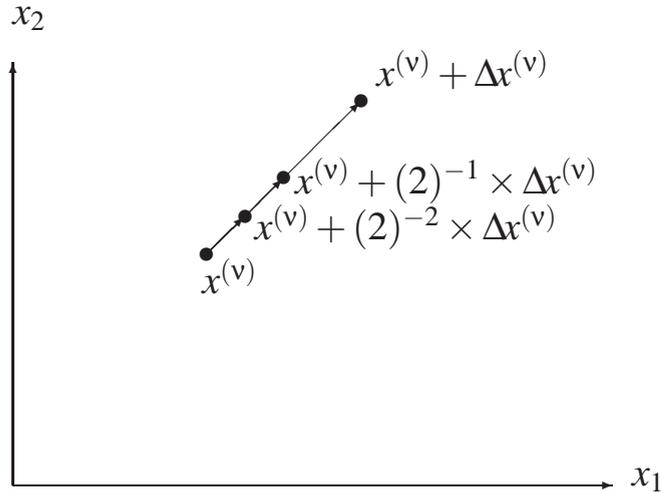
$x^{(\nu)}$

x_1

Fig. 7.10. Illustration of back-tracking in Armijo step-size rule.

7.4.2.4 Example

- Again consider the arctan function and $x^{(\nu)} = 5$.

$$\begin{aligned}
\Delta x^{(\nu)} &= -\left[J(x^{(\nu)})\right]^{-1} g(x^{(\nu)}), \\
&\approx -35.7.
\end{aligned}$$

- Set $\delta = 0.5$.
- The dotted lines in Figure 7.11 bound the set of points of the form $\begin{bmatrix} x^{(\nu)} + \alpha^{(\nu)} \Delta x^{(\nu)} \\ \gamma \end{bmatrix}$ satisfying:

$$0 \leq \alpha^{(\nu)} \leq 1,$$
$$-(1 - \delta \alpha^{(\nu)}) \left\| g(x^{(\nu)}) \right\| \leq \gamma \leq (1 - \delta \alpha^{(\nu)}) \left\| g(x^{(\nu)}) \right\|.$$
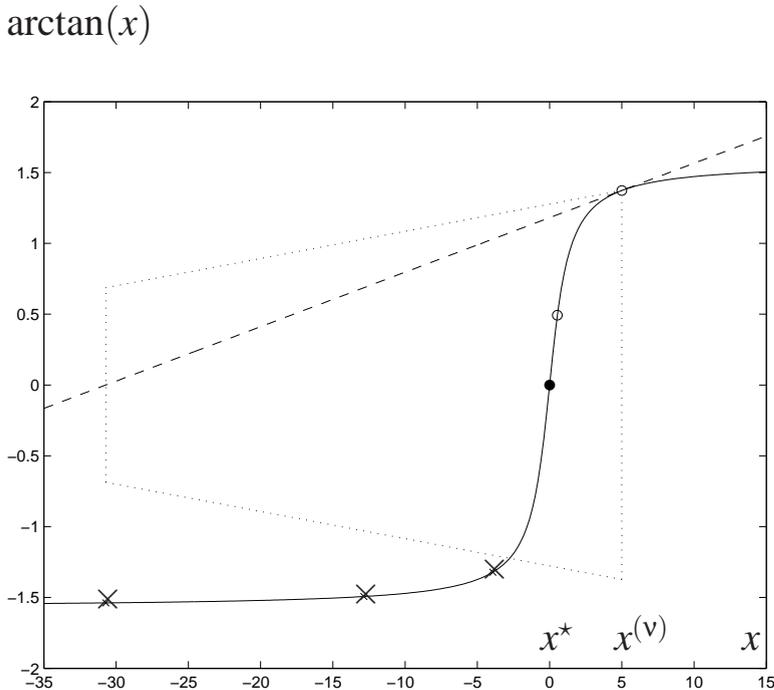
Example, continued

arctan(x)

Fig. 7.11. Armijo update applied to solving equation with arctan function (shown solid). The first-order Taylor approximation about $x^{(\nu)} = 5$ is shown dashed. The point $\begin{bmatrix} x^{(\nu)} \\ g(x^{(\nu)}) \end{bmatrix}$ is illustrated by the rightmost $\circ$, while the solution to the equation $g(x) = \mathbf{0}$ is shown with a $\bullet$. The dotted lines bound the region of acceptance for the Armijo rule with $\delta = 0.5$. The leftmost three $\times$ do not satisfy the Armijo rule. The updated iterate is illustrated by the leftmost $\circ$.

- Using step-sizes of the form (7.30) results in tentative updated iterates and corresponding function values of:

$$
\begin{aligned}
x^{(\nu)} + \Delta x^{(\nu)} &\approx -30.7, & g(x^{(\nu)} + \Delta x^{(\nu)}) &\approx -1.54, \\
x^{(\nu)} + (2)^{-1} \times \Delta x^{(\nu)} &\approx -12.9, & g(x^{(\nu)} + (2)^{-1} \times \Delta x^{(\nu)}) &\approx -1.49, \\
x^{(\nu)} + (2)^{-2} \times \Delta x^{(\nu)} &\approx -3.93, & g(x^{(\nu)} + (2)^{-2} \times \Delta x^{(\nu)}) &\approx -1.32, \\
x^{(\nu)} + (2)^{-3} \times \Delta x^{(\nu)} &\approx 0.54, & g(x^{(\nu)} + (2)^{-3} \times \Delta x^{(\nu)}) &\approx 0.49.
\end{aligned}
$$

7.4.2.5 *Choice of* δ

- If the parameter δ is close to one then it may take many reductions of $\alpha^{(v)}$ to satisfy (7.29).
- δ is often chosen to be considerably less than one.

7.4.2.6 *Variations*

- There are other variations on (7.29)–(7.30) that seek to avoid unnecessary "back-tracking."

7.4.2.7 *Discussion*

- Step-size rules can significantly aid in convergence from an initial guess that is far from the solution.

7.4.3 *Computational effort*

- Variations on the Newton–Raphson method that require less effort per iteration will tend to perform better overall than the exact Newton–Raphson method.

7.5 Sensitivity and large change analysis

7.5.1 Sensitivity

7.5.1.1 Implicit function theorem

Corollary 7.5 *Let $g : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^n$ be partially differentiable with continuous partial derivatives. Consider solutions of the equations $g(x; \chi) = \mathbf{0}$, where χ is a parameter. Suppose that $x^\star$ satisfies:*

$$g(x^\star; \mathbf{0}) = \mathbf{0}.$$

We call $x = x^\star$ the base-case solution and $\chi = \mathbf{0}$ the base-case parameters. Define the (parameterized) Jacobian $J : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^{n \times n}$ by:

$$\forall x \in \mathbb{R}^n, \forall \chi \in \mathbb{R}^s, J(x; \chi) = \frac{\partial g}{\partial x}(x; \chi).$$

Suppose that $J(x^\star; \mathbf{0})$ is non-singular. Then, there is a solution to $g(x; \chi) = \mathbf{0}$ for χ in a neighborhood of the base-case values of the parameters $\chi = \mathbf{0}$. The sensitivity of the solution $x^\star$ to variation of the

parameters χ, evaluated at the base-case $\chi = \mathbf{0}$, is given by:

$$\frac{\partial x^\star}{\partial \chi}(\mathbf{0}) = -[J(x^\star; \mathbf{0})]^{-1} K(x^\star; \mathbf{0}),$$

where $K : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^{n \times s}$ is defined by:

$$\forall x \in \mathbb{R}^n, \forall \chi \in \mathbb{R}^s, K(x; \chi) = \frac{\partial g}{\partial \chi}(x; \chi).$$

$\square$

- If $J(x^\star; \mathbf{0})$ has already been factorized then the calculation of the sensitivity requires one forwards and backwards substitution for each entry of χ.

7.5.1.2 Example

- Suppose that $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is defined by:

$$\forall x \in \mathbb{R}, \forall \chi \in \mathbb{R}, g(x; \chi) = (x - 2 - \sin \chi)^3 + 1.$$

- The base-case solution is $x^\star = 1$.
- We consider the sensitivity of the solution to the parameter χ, evaluated at $\chi = 0$.

Example, continued

- Using Corollary 7.5, we have that the sensitivity is given by:

$$-[J(x^\star;\mathbf{0})]^{-1}K(x^\star;\mathbf{0}),$$

- where $J : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^{n \times n}$ and $K : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^{n \times s}$ are defined by:

$$\begin{aligned}
\forall x \in \mathbb{R}^n, \forall \chi \in \mathbb{R}^s, J(x;\chi) &= \frac{\partial g}{\partial x}(x;\chi), \\
&= 3(x - 2 - \sin\chi)^2, \\
J(x^\star;\mathbf{0}) &= 3, \\
\forall x \in \mathbb{R}^n, \forall \chi \in \mathbb{R}^s, K(x;\chi) &= \frac{\partial g}{\partial \chi}(x;\chi), \\
&= 3(x - 2 - \sin\chi)^2(-\cos\chi), \\
K(x^\star;\mathbf{0}) &= -3.
\end{aligned}$$

- Substituting, the sensitivity is 1.

7.5.2 *Large changes*

- Use the iterative techniques we have developed, using as initial guess the solution to the base-case.

7.6 Summary

- The Newton–Raphson method and variants,
- Local convergence results,
- Globalization procedures,
- Sensitivity analysis.

8

Solution of the non-linear simultaneous equations case studies

- Non-linear DC circuit in Section 8.1, and
- Power flow problem in Section 8.2.

8.1 Analysis of a non-linear direct current circuit

- The circuit satisfies $g(x) = \mathbf{0}$, where $g : \mathbb{R}^4 \to \mathbb{R}^4$ was defined in (6.6):

$$\forall x \in \mathbb{R}^4, g(x) = \begin{bmatrix} \left(\frac{1}{R_a}\right) x_1 + i_b(x_1 - x_2) - I_1 \\ -i_b(x_1 - x_2) + \left(\frac{1}{R_c} + \frac{1}{R_d}\right) x_2 + \left(-\frac{1}{R_d}\right) x_3 \\ \left(-\frac{1}{R_d}\right) x_2 + \left(\frac{1}{R_d} + \frac{1}{R_e} + \frac{1}{R_f}\right) x_3 + \left(-\frac{1}{R_f}\right) x_4 \\ \left(-\frac{1}{R_f}\right) x_3 + \left(\frac{1}{R_f}\right) x_4 + i_g(x_4) \end{bmatrix}.$$
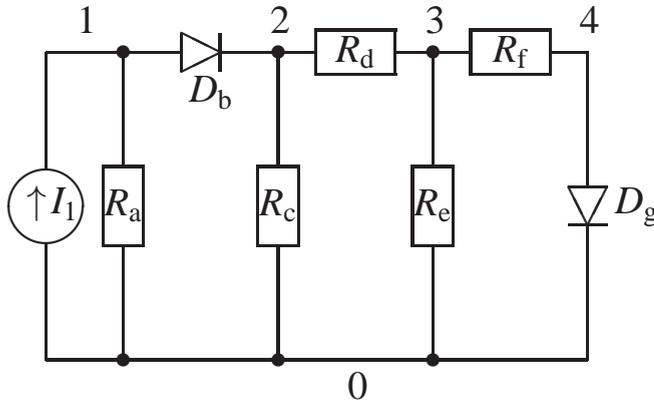

Fig. 8.1. The non-linear DC circuit from Figure 6.3.

8.1.1 Jacobian

$$\forall x \in \mathbb{R}^4, J(x) =$$

$$\begin{bmatrix} \left(\frac{1}{R_a}\right) + \frac{di_b}{dV_b}(x_1 - x_2) & -\frac{di_b}{dV_b}(x_1 - x_2) & 0 & 0 \\ -\frac{di_b}{dV_b}(x_1 - x_2) & \frac{di_b}{dV_b}(x_1 - x_2) + \left(\frac{1}{R_c} + \frac{1}{R_d}\right) & \left(-\frac{1}{R_d}\right) & 0 \\ 0 & \left(-\frac{1}{R_d}\right) & \left(\frac{1}{R_d} + \frac{1}{R_e} + \frac{1}{R_f}\right) & \left(-\frac{1}{R_f}\right) \\ 0 & 0 & \left(-\frac{1}{R_f}\right) & \left(\frac{1}{R_f}\right) + \frac{di_g}{dV_g}(x_4) \end{bmatrix}.$$

(8.1)

- The Jacobian is similar in appearance to the admittance matrix for a linear circuit, with the same sort of sparsity structure:
  - non-zeros on the diagonals, and
  - non-zeros on the off-diagonals corresponding to branches.
- For the diodes, we have **incremental** admittances evaluated at x instead of admittances.

8.1.2 Initial guess

- In the absence of a better guess, $x^{(0)} = \mathbf{0}$ may be a reasonable initial guess for our circuit.
- Better guesses will save on computation time and occasionally make the difference between successful and unsuccessful application of the algorithm.

8.1.3 Calculation of iterates

$$
\begin{aligned}
J(x^{(0)})\Delta x^{(0)} &= -g(x^{(0)}), \\
x^{(1)} &= x^{(0)} + \Delta x^{(0)}.
\end{aligned}
$$

8.1.4 *Application of chord and Kantorovich theorems*

- Applying the chord and Kantorovich theorems can require considerable effort even for simple problems.
- The theorems will run into difficulty if the entries in the Jacobian vary greatly with their argument because this will cause a large value for the Lipschitz constant c.
- Large variation of the entries in the Jacobian occurs in the diode model and other models with cut-off/cut-on characteristics where the slope of the current versus voltage characteristic varies from near zero to very large.

- We can find that:

$$\left\| g(x^{(1)}) \right\|_2 > \left\| g(x^{(0)}) \right\|_2.$$

- The chord and Kantorovich convergence theorems we have presented are *local* in nature.
- Their conclusions do not help us if we are solving a circuit for the first time and do not know which diodes will be conducting and which will be off.

8.1.5 Step-size rules

- A step-size rule can significantly aid in convergence even when the Jacobian varies greatly.
- The Armijo rule will guarantee that $\left\|g(x^{(1)})\right\|_2 < \left\|g(x^{(0)})\right\|_2$ and improve convergence.

8.1.6 Stopping criteria

- If the measurement is accurate to, say, 0.1%, then it is superfluous to try to solve the equations to *far* better than this accuracy.

- If all measurements were accurate to around $0.1\% = 10^{-3}$, a suitable stopping criterion would be:

$$\left\| g(x^{(\nu)}) \right\|_\infty \le 10^{-4}, \text{ and}$$

$$\left\| \Delta x^{(\nu-1)} \right\|_2 \le 10^{-4} \text{ or } \left\| \Delta x^{(\nu-1)} \right\|_\infty \le 10^{-4}.$$

- We might require that this condition be satisfied over several successive iterates.

- We can also try to apply the chord and Kantorovich theorems to the current iterate, $x^{(\nu)}$ say, re-interpreted as a new initial guess.

8.1.7 Circuit changes

- Now we suppose that the equations are *parameterized* by a parameter $\chi \in \mathbb{R}^s$.
- That is, $g : \mathbb{R}^4 \times \mathbb{R}^s \to \mathbb{R}^4$, with the base-case solution corresponding to $\chi = \mathbf{0}$.

8.1.7.1 Sensitivity

- Sensitivity of the base-case solution to changes in χ:

$$\frac{\partial x^\star}{\partial \chi}(\mathbf{0}) = -[J(x^\star;\mathbf{0})]^{-1} K(x^\star;\mathbf{0}),$$

- where $J : \mathbb{R}^4 \times \mathbb{R}^s \to \mathbb{R}^{4 \times 4}$ and $K : \mathbb{R}^4 \times \mathbb{R}^s \to \mathbb{R}^{4 \times s}$ are defined by:

$$\forall x \in \mathbb{R}^4, \forall \chi \in \mathbb{R}^s, J(x;\chi) = \frac{\partial g}{\partial x}(x;\chi), K(x;\chi) = \frac{\partial g}{\partial \chi}(x;\chi).$$

8.1.7.2 *Large change analysis*

- Apply the Newton–Raphson method (or one of the variants) to the changed system using an initial guess for the changed system that is given by the base-case solution $x^\star$ or by an estimate of the change-case solution using sensitivity analysis.
- For a change in a resistor or diode, we can also update the Jacobian using a rank one update.
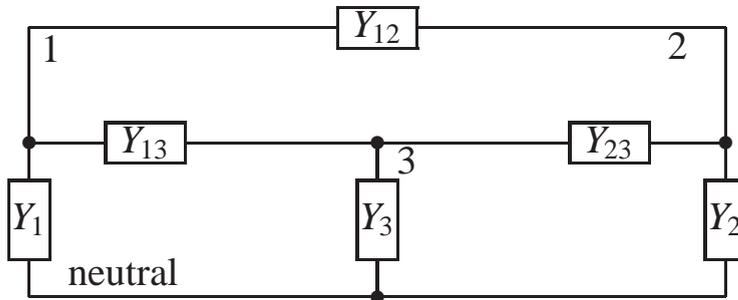
8.2 Analysis of an electric power system

Fig. 8.2. Per-phase equivalent circuit model repeated from Figure 6.10.

8.2.1 Jacobian

8.2.1.1 Terms

- The entries in $g : \mathbb{R}^n \to \mathbb{R}^n$ are either of the form $p_\ell : \mathbb{R}^n \to \mathbb{R}$:

$$\forall x \in \mathbb{R}^n, p_\ell(x) = \sum_{k \in \mathbb{J}(\ell) \cup \{\ell\}} u_\ell u_k [G_{\ell k} \cos(\theta_\ell - \theta_k) + B_{\ell k} \sin(\theta_\ell - \theta_k)] - P_\ell,$$

- or of the form $q_\ell : \mathbb{R}^n \to \mathbb{R}$:

$$\forall x \in \mathbb{R}^n, q_\ell(x) = \sum_{k \in \mathbb{J}(\ell) \cup \{\ell\}} u_\ell u_k [G_{\ell k} \sin(\theta_\ell - \theta_k) - B_{\ell k} \cos(\theta_\ell - \theta_k)] - Q_\ell.$$

- The entries in the vector x are either of the form θ_k or of the form u_k.

- Four qualitative types of partial derivative terms corresponding to each combination:

$$\forall x \in \mathbb{R}^n, \frac{\partial p_\ell}{\partial \theta_k}(x)$$

$$= \begin{cases} \displaystyle\sum_{j \in \mathbb{J}(\ell)} u_\ell u_j[-G_{\ell j}\sin(\theta_\ell - \theta_j) + B_{\ell j}\cos(\theta_\ell - \theta_j)], & \text{if } k = \ell, \\[2ex] u_\ell u_k[G_{\ell k}\sin(\theta_\ell - \theta_k) - B_{\ell k}\cos(\theta_\ell - \theta_k)], & \text{if } k \in \mathbb{J}(\ell), \\[1ex] 0, & \text{otherwise,} \end{cases}$$

$$\forall x \in \mathbb{R}^n, \frac{\partial p_\ell}{\partial u_k}(x)$$

$$= \begin{cases} 2u_\ell G_{\ell\ell} + \displaystyle\sum_{j \in \mathbb{J}(\ell)} u_j[G_{\ell j}\cos(\theta_\ell - \theta_j) + B_{\ell j}\sin(\theta_\ell - \theta_j)], & \text{if } k = \ell, \\[2ex] u_\ell[G_{\ell k}\cos(\theta_\ell - \theta_k) + B_{\ell k}\sin(\theta_\ell - \theta_k)], & \text{if } k \in \mathbb{J}(\ell), \\[1ex] 0, & \text{otherwise,} \end{cases}$$

$$\forall x \in \mathbb{R}^n, \frac{\partial q_\ell}{\partial \theta_k}(x)$$

$$= \begin{cases} \displaystyle\sum_{j \in \mathbb{J}(\ell)} u_\ell u_j [G_{\ell j} \cos(\theta_\ell - \theta_j) + B_{\ell j} \sin(\theta_\ell - \theta_j)], & \text{if } k = \ell, \\ u_\ell u_k [-G_{\ell k} \cos(\theta_\ell - \theta_k) - B_{\ell k} \sin(\theta_\ell - \theta_k)], & \text{if } k \in \mathbb{J}(\ell), \\ 0, & \text{otherwise}, \end{cases}$$

$$\forall x \in \mathbb{R}^n, \frac{\partial q_\ell}{\partial u_k}(x)$$

$$= \begin{cases} -2u_\ell B_{\ell\ell} + \displaystyle\sum_{j \in \mathbb{J}(\ell)} u_j [G_{\ell j} \sin(\theta_\ell - \theta_j) - B_{\ell j} \cos(\theta_\ell - \theta_j)], & \text{if } k = \ell, \\ u_\ell [G_{\ell k} \sin(\theta_\ell - \theta_k) - B_{\ell k} \cos(\theta_\ell - \theta_k)], & \text{if } k \in \mathbb{J}(\ell), \\ 0, & \text{otherwise}. \end{cases}$$

8.2.1.2 Partitioning by types of terms

- Order the entries in g so that all the equations for real power appear first in a sub-vector p followed by all the equations for reactive power in a sub-vector q.
- Partition x so that all the voltage angles appear first in a sub-vector θ followed by all the voltage magnitudes in a sub-vector u.

- We can partition the Jacobian into four blocks:

$$\forall x \in \mathbb{R}^n, J(x) = \begin{bmatrix} J_{p\theta}(x) & J_{pu}(x) \\ J_{q\theta}(x) & J_{qu}(x) \end{bmatrix}, \tag{8.2}$$

$$\forall x \in \mathbb{R}^n, J_{p\theta}(x) = \frac{\partial p}{\partial \theta}(x),$$

$$\forall x \in \mathbb{R}^n, J_{pu}(x) = \frac{\partial p}{\partial u}(x),$$

$$\forall x \in \mathbb{R}^n, J_{q\theta}(x) = \frac{\partial q}{\partial \theta}(x),$$

$$\forall x \in \mathbb{R}^n, J_{qu}(x) = \frac{\partial q}{\partial u}(x).$$

8.2.1.3 *Sparsity*

- Each of the four blocks in (8.3) has the same sparsity structure as the bus admittance matrix.

8.2.1.4 Symmetry

- The blocks $J_{p\theta}, J_{pu}, J_{q\theta}$, and J_{qu} are not symmetric and $[J_{pu}]^{\dagger} \neq J_{q\theta}$.
- That is, the Jacobian as a whole is not symmetric.

8.2.1.5 Partitioning by bus number

- An alternative to partitioning by the types of terms is to partition the Jacobian into blocks based on the bus number.
- As discussed in Section 5.5.4.2, we can treat each 2×2 block as a single "entry" in our sparse matrix.
- We can use block pivoting as discussed in Section 5.5.4.2.
- We can treat each 2×2 block as a single entity in factorization by explicitly inverting the block using the formula for the inverse of a 2×2 matrix.
- We will not use this approach for solving the power flow problem.
- In some extensions of this problem block pivoting can be exploited to speed up calculations considerably.

8.2.2 Initial guess

- A sensible choice for the initial guess for the voltage magnitude is $u^{(0)} = \mathbf{1}$, where $\mathbf{1}$ is the vector of all ones.
- A possible guess for the voltage angle is $\theta^{(0)} = \mathbf{0}$.
- These choices of initial guess for voltage angle and magnitude are called a "flat start."

8.2.3 Calculation of iterates

$$\begin{bmatrix} J_{p\theta}(x) & J_{pu}(x) \\ J_{q\theta}(x) & J_{qu}(x) \end{bmatrix} \begin{bmatrix} \Delta\theta^{(v)} \\ \Delta u^{(v)} \end{bmatrix} = -\begin{bmatrix} p(x^{(v)}) \\ q(x^{(v)}) \end{bmatrix}, \tag{8.3}$$

$$\theta^{(v+1)} = \theta^{(v)} + \Delta\theta^{(v)}, \tag{8.4}$$

$$u^{(v+1)} = u^{(v)} + \Delta u^{(v)}. \tag{8.5}$$

8.2.4 Approximation of the Jacobian and update

8.2.4.1 Chord and Shamanskii updates

- Using a flat start, $x^{(0)} = \begin{bmatrix} \theta^{(0)} \\ u^{(0)} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix}$, as our initial guess, the entries for the Jacobian become:

$$\frac{\partial p_\ell}{\partial \theta_k}(x^{(0)}) = \begin{cases} \displaystyle\sum_{j \in \mathbb{J}(\ell)} B_{\ell j}, & \text{if } k = \ell, \\ -B_{\ell k}, & \text{if } k \in \mathbb{J}(\ell), \\ 0, & \text{otherwise}, \end{cases}$$

$$\frac{\partial p_\ell}{\partial u_k}(x^{(0)}) = \begin{cases} 2G_{\ell\ell} + \displaystyle\sum_{j \in \mathbb{J}(\ell)} G_{\ell j}, & \text{if } k = \ell, \\ G_{\ell k}, & \text{if } k \in \mathbb{J}(\ell), \\ 0, & \text{otherwise}, \end{cases}$$

$$\frac{\partial q_\ell}{\partial \theta_k}(x^{(0)}) = \begin{cases} \sum\limits_{j \in \mathbb{J}(\ell)} G_{\ell j}, & \text{if } k = \ell, \\ -G_{\ell k}, & \text{if } k \in \mathbb{J}(\ell), \\ 0, & \text{otherwise,} \end{cases}$$

$$\frac{\partial q_\ell}{\partial u_k}(x^{(0)}) = \begin{cases} -2B_{\ell\ell} - \sum\limits_{j \in \mathbb{J}(\ell)} B_{\ell j}, & \text{if } k = \ell, \\ -B_{\ell k}, & \text{if } k \in \mathbb{J}(\ell), \\ 0, & \text{otherwise.} \end{cases}$$

8.2.4.2 *Approximating particular terms*

- We will first approximate the Jacobian by:

    (i) neglecting all the terms in the blocks J_{pu} and $J_{q\theta}$, and
    (ii) approximating some of the terms in the blocks $J_{p\theta}$ and J_{qu}.

- Neglecting terms in the blocks increases the sparsity of the equations.
- Approximations to the terms in $J_{p\theta}$ and J_{qu} then yield a linear system that is similar to the Jacobian used in the chord update with a flat start.

Neglecting terms

- As noted in Section 6.2.4.4, typically:

$$|G_{\ell k}| \ll |B_{\ell k}|. \tag{8.6}$$

- A typical limit on angle differences is $|\theta_\ell - \theta_k| \leq \frac{\pi}{4}$.

$$
\begin{aligned}
|\sin(\theta_\ell - \theta_k)| &\approx |\theta_\ell - \theta_k|, \text{ for small angle differences in radians,} \\
&\ll 1, \text{ for small angle differences,} \tag{8.7} \\
\cos(\theta_\ell - \theta_k) &\approx 1, \text{ for small angle differences,} \tag{8.8} \\
u_\ell &\approx 1. \tag{8.9}
\end{aligned}
$$

Neglecting terms, continued

$$\frac{\partial p_\ell}{\partial \theta_k}(x) = u_\ell u_k [G_{\ell k} \sin(\theta_\ell - \theta_k) - B_{\ell k} \cos(\theta_\ell - \theta_k)],$$

$$\approx G_{\ell k}(\theta_\ell - \theta_k) - B_{\ell k},$$
$$\text{since } u_\ell \approx 1, u_k \approx 1, \cos(\theta_\ell - \theta_k) \approx 1,$$
$$\approx -B_{\ell k}, \text{ since } |\theta_\ell - \theta_k| \ll 1 \text{ and } |G_{\ell k}| \ll |B_{\ell k}|, \qquad (8.10)$$

$$\left|\frac{\partial p_\ell}{\partial u_k}(x)\right| = \left|u_\ell [G_{\ell k} \cos(\theta_\ell - \theta_k) + B_{\ell k} \sin(\theta_\ell - \theta_k)]\right|,$$

$$\approx |G_{\ell k} + B_{\ell k}(\theta_\ell - \theta_k)|,$$
$$\text{since } u_\ell \approx 1, \cos(\theta_\ell - \theta_k) \approx 1, \sin(\theta_\ell - \theta_k) \approx (\theta_\ell - \theta_k),$$
$$\ll |B_{\ell k}|, \text{ since } |\theta_\ell - \theta_k| \ll 1 \text{ and } |G_{\ell k}| \ll |B_{\ell k}|, \qquad (8.11)$$

Neglecting terms, continued

$$\left|\frac{\partial q_\ell}{\partial \theta_k}(x)\right| = \left|u_\ell u_k[-G_{\ell k}\cos(\theta_\ell - \theta_k) - B_{\ell k}\sin(\theta_\ell - \theta_k)]\right|,$$

$$\approx \left|-G_{\ell k} - B_{\ell k}(\theta_\ell - \theta_k)\right|,$$

$$\text{since } u_\ell \approx 1, u_k \approx 1, \cos(\theta_\ell - \theta_k) \approx 1, \sin(\theta_\ell - \theta_k) \approx \theta_\ell - \theta_k,$$

$$\ll |B_{\ell k}|, \text{ since } |\theta_\ell - \theta_k| \ll 1 \text{ and } |G_{\ell k}| \ll |B_{\ell k}|, \tag{8.12}$$

$$\frac{\partial q_\ell}{\partial u_k}(x) = u_\ell[G_{\ell k}\sin(\theta_\ell - \theta_k) - B_{\ell k}\cos(\theta_\ell - \theta_k)],$$

$$\approx G_{\ell k}(\theta_\ell - \theta_k) - B_{\ell k},$$

$$\text{since } u_\ell \approx 1, \cos(\theta_\ell - \theta_k) \approx 1, \sin(\theta_\ell - \theta_k) \approx (\theta_\ell - \theta_k),$$

$$\approx -B_{\ell k}, \text{ since } |\theta_\ell - \theta_k| \ll 1 \text{ and } |G_{\ell k}| \ll |B_{\ell k}|. \tag{8.13}$$

- These approximations reflect the qualitative observation that real power flow is mostly determined by differences in voltage angles across lines, while reactive power flow is mostly determined by voltage magnitude differences

Neglecting terms

- If we neglect all the terms in J_{pu} and $J_{q\theta}$, then we can then approximate the Jacobian by $J(x) \approx \begin{bmatrix} J_{p\theta}(x) & \mathbf{0} \\ \mathbf{0} & J_{qu}(x) \end{bmatrix}$.

$$\begin{bmatrix} J_{p\theta}(x) & \mathbf{0} \\ \mathbf{0} & J_{qu}(x) \end{bmatrix} \begin{bmatrix} \Delta\theta^{(v)} \\ \Delta u^{(v)} \end{bmatrix} = -\begin{bmatrix} p(x^{(v)}) \\ q(x^{(v)}) \end{bmatrix}, \tag{8.14}$$

$$\theta^{(v+1)} = \theta^{(v)} + \Delta\theta^{(v)},$$
$$u^{(v+1)} = u^{(v)} + \Delta u^{(v)}.$$

- These are called the **decoupled** Newton–Raphson update equations:

$$J_{p\theta}(x)\Delta\theta^{(v)} = -p(x^{(v)}), \tag{8.15}$$
$$J_{qu}(x)\Delta u^{(v)} = -q(x^{(v)}). \tag{8.16}$$

Approximating terms

- In addition to assuming that $|G_{\ell k}| \ll |B_{\ell k}|$ and that $\cos(\theta_\ell - \theta_k) \approx 1$, we will assume that:

  (i) for any bus ℓ, the magnitude of the voltages u_j at buses $j \in \mathbb{J}(\ell)$ is approximately the same as the magnitude of the voltage u_ℓ at ℓ, and

  (ii) $B_{\ell\ell} \approx -\sum_{j \in \mathbb{J}(\ell)} B_{\ell j}$.

$$
\begin{aligned}
\frac{\partial p_\ell}{\partial \theta_\ell}(x) \; &= \; \sum_{j \in \mathbb{J}(\ell)} u_\ell u_j [-G_{\ell j}\sin(\theta_\ell - \theta_j) + B_{\ell j}\cos(\theta_\ell - \theta_j)], \\
&\approx \; \sum_{j \in \mathbb{J}(\ell)} (u_\ell)^2 [-G_{\ell j}\sin(\theta_\ell - \theta_j) + B_{\ell j}\cos(\theta_\ell - \theta_j)], \\
&\qquad\qquad \text{assuming } u_j \approx u_\ell \text{ for } j \in \mathbb{J}(\ell), \\
&\approx \; \sum_{j \in \mathbb{J}(\ell)} (u_\ell)^2 B_{\ell j}, \text{ since } |G_{\ell k}| \ll |B_{\ell k}| \text{ and } \cos(\theta_\ell - \theta_k) \approx 1, \\
&\approx \; -(u_\ell)^2 B_{\ell\ell}, \text{ since } B_{\ell\ell} \approx -\sum_{j \in \mathbb{J}(\ell)} B_{\ell j}.
\end{aligned}
$$

Approximating terms, continued

$$\frac{\partial q_\ell}{\partial u_\ell}(x) \; = \; -2u_\ell B_{\ell\ell} + \sum_{j \in \mathbb{J}(\ell)} u_j [G_{\ell j} \sin(\theta_\ell - \theta_j) - B_{\ell j} \cos(\theta_\ell - \theta_j)],$$

$$\approx \; -2u_\ell B_{\ell\ell} + \sum_{j \in \mathbb{J}(\ell)} u_\ell [G_{\ell j} \sin(\theta_\ell - \theta_j) - B_{\ell j} \cos(\theta_\ell - \theta_j)],$$

$$\text{assuming } u_j \approx u_\ell \text{ for } j \in \mathbb{J}(\ell),$$

$$\approx \; -2u_\ell B_{\ell\ell} - \sum_{j \in \mathbb{J}(\ell)} u_\ell B_{\ell j}, \text{ since } |G_{\ell k}| \ll |B_{\ell k}| \text{ and } \cos(\theta_\ell - \theta_k) \approx 1,$$

$$\approx \; -u_\ell B_{\ell\ell}, \text{ assuming } B_{\ell\ell} \approx -\textstyle\sum_{j \in \mathbb{J}(\ell)} B_{\ell j},$$

$$\frac{\partial p_\ell}{\partial \theta_k}(x) \; \approx \; -u_\ell B_{\ell k} u_k,$$

$$\frac{\partial q_\ell}{\partial u_k}(x) \; \approx \; -u_\ell B_{\ell k}.$$

Approximating terms, continued

$$\frac{\partial p_\ell}{\partial \theta_k}(x) = 0,$$

$$= -u_\ell B_{\ell k} u_k,$$

$$\frac{\partial q_\ell}{\partial u_k}(x) = 0,$$

$$= -u_\ell B_{\ell k}.$$

- In summary, the approximations $\frac{\partial p_\ell}{\partial \theta_k}(x) \approx -u_\ell B_{\ell k} u_k$ and

$\frac{\partial q_\ell}{\partial u_k}(x) \approx -u_\ell B_{\ell k}$ apply for all ℓ and k.

Compact representation

- Define the matrix B to be the imaginary part of the bus admittance matrix A.
- Define U to be the diagonal matrix having diagonal entries equal to the corresponding entries of u.

$$J_{p\theta}(x) \approx -UBU,$$
$$J_{qu}(x) \approx -UB.$$

At iteration ν, the decoupled equations (8.14) can therefore be approximated by:

$$-U^{(\nu)}BU^{(\nu)}\Delta\theta^{(\nu)} = -p(x^{(\nu)}), \qquad (8.17)$$
$$-U^{(\nu)}B\Delta u^{(\nu)} = -q(x^{(\nu)}). \qquad (8.18)$$

Pre-conditioning and scaling variables

- By moving $u^{(\nu)}$ to the right-hand sides of (8.17) and (8.18) and defining $\Delta\phi^{(\nu)} = u^{(\nu)}\Delta\theta^{(\nu)}$, we obtain the equivalent system:

$$-B\Delta\phi^{(\nu)} = -[U^{(\nu)}]^{-1} p(x^{(\nu)}), \qquad (8.19)$$

$$-B\Delta u^{(\nu)} = -[U^{(\nu)}]^{-1} q(x^{(\nu)}). \qquad (8.20)$$

- The coefficient matrix $(-B)$ on the left-hand sides of both (8.19) and (8.20) is constant and symmetric.
- To solve (8.19) and (8.20), we need only perform *LU* factorization of $(-B)$ once, not once per iteration.
- Once $\Delta\phi^{(\nu)}$ is known, $\Delta\theta^{(\nu)}$ can be calculated using:

$$\Delta\theta^{(\nu)} = [U^{(\nu)}]^{-1}\Delta\phi^{(\nu)}. \qquad (8.21)$$

Discussion

- The advantage of using a constant coefficient matrix in (8.19) and (8.20) is that it significantly reduces the computational effort per iteration.
- The approximations we have described are not always very good.
- But found to work in practice to decrease computational effort overall.

8.2.4.3 Quasi-Newton methods

- Quasi-Newton methods can also be applied to solve the equations.
- Equations (8.19) and (8.20) specify a suitable initialization for the approximation to the Jacobian.

8.2.4.4 Iterative methods

- Instead of directly solving the linear equations for the Newton–Raphson update, it is also possible to use an iterative algorithm, such as the conjugate gradient method.

8.2.5 Step-size rules

- A step-size rule can aid in convergence.

8.2.6 Stopping criteria

- Require a sufficiently small value of the norm of the:
  - change between successive iterates, and
  - deviation of the entries of g from zero.

8.2.7 *Circuit changes*

8.2.7.1 *Sensitivity*

$$\frac{\partial x^\star}{\partial \chi}(\mathbf{0}) = -[J(x^\star;\mathbf{0})]^{-1}K(x^\star;\mathbf{0}),$$

$$\forall x \in \mathbb{R}^n, \forall \chi \in \mathbb{R}^s, J(x;\chi) = \frac{\partial g}{\partial x}(x;\chi),$$

$$\forall x \in \mathbb{R}^n, \forall \chi \in \mathbb{R}^s, K(x;\chi) = \frac{\partial g}{\partial \chi}(x;\chi).$$

8.2.7.2 *Large change analysis*

- Large changes to the real and reactive injections into the system can be analyzed by restarting the Newton–Raphson updates based on the solution to the base-case system.
- If the fast decoupled update equations are used, no changes are necessary to the Jacobian.
- Changes to the transmission lines require an update to the Jacobian even if the approximate Jacobian is used.