

Chapter 7

American Option Pricing

In contrast to European Options, an American Option can be exercised at **any** time $t \leq T$. This means that the payoff function is (S denotes again the price of the underlying asset, K is the exercise price, i.e. the strike)

$$\begin{aligned} V_C^{\text{am}}(S, t) &= (S_t - K)^+ && \text{for a call and} \\ V_P^{\text{am}}(S, t) &= (K - S_t)^+ && \text{for a put.} \end{aligned} \tag{7.0.1}$$

This means, at expiration time T , the payoff coincides with European Options. Under no-arbitrage assumptions, one can easily show the following a-priori bounds

$$\begin{aligned} V_P^{\text{am}}(S, t) &\geq (K - S)^+ \\ V_C^{\text{am}}(S, t) &\geq (S - K)^+ \end{aligned} \quad \forall S, t$$

and trivially

$$V^{\text{am}} \geq V^{\text{eur}} ,$$

since a European Option is a special case of an American Option.

Again, we do not go into details of the modelling of American options but refer to any lecture on Financial Mathematics and the literature, e.g. [4, 5] which are also the basis for this chapter. Here we concentrate only on those facts that are relevant for the numerical simulation of these financial processes.

The *contact point* S_f is defined as follows

$$V_P^{\text{am}}(S_f, t) = K - S_f \quad (0 < S_f < K) . \tag{7.0.2}$$

Note that S_f depends on the time t , i.e., $S_f = S_f(t)$. The contact point $S_f(t)$ can be characterized by

$$\begin{cases} V_P^{\text{am}}(S, t) > (K - S)^+ & \text{for } S > S_f(t) , \\ V_P^{\text{am}}(S, t) = K - S & \text{for } S \leq S_f(t) . \end{cases} \quad (7.0.3)$$

The location of the manifold $S_f(t)$, $t \in (0, T)$ is *unknown* a priori. Since this graph is the interface between the ‘exercise’ and the ‘no-exercise’ region (see below), we are faced with a so called *free boundary-value problem*. These kind of problems also occur in several other application, e.g., the propagation of waves in a medium (water waves, acoustic waves), plastic deformation processes and so on. Most of what is said in this chapter can also be applied to this kind of ‘industrial’ problems

The interpretation of the contact point for a put is as follows. The holder should exercise as soon as the price of the asset reaches $S_f(t)$. The corresponding time instant t_S is called *stopping time*. For a filtration \mathcal{F}_t , a random variable τ that is \mathcal{F}_t -measurable for all $t \geq 0$ is called *stopping time*.

Boundary conditions: The slope $\frac{\partial V}{\partial S}$ with which $V_P^{\text{am}}(S, t)$ touches the straight line $K - S$ at $S_f(t)$ is used as a boundary condition. Note that $K - S$ has the slope $-1 = \frac{\partial}{\partial S}(K - S)$, hence we require

$$\frac{\partial}{\partial S} V_P^{\text{am}}(S_f(t), t) = -1, \quad (7.0.4)$$

which results in a tangential touching. This is the so-called *high contact condition*.

For the (somewhat hypothetical) case of a *perpetual option* (i.e., for maturity $T = \infty$), we obtain an asymptotic condition (that can be calculated analytically). We will come to this point later.

In general, we obtain two boundary conditions, namely (7.0.2), (7.0.4).

For the American call, we need to include also dividend yields δ (otherwise they coincide with a European call option), i.e., the Black-Scholes equation takes the form

$$\frac{\partial}{\partial t} V(S, t) + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2}{\partial S^2} V(S, t) + (r - \delta) S \frac{\partial}{\partial S} V(S, t) - rV(S, t) = 0 \quad (7.0.5)$$

In fact, one can show that if $\delta = 0$ an early exercise does not pay off for a call. Since $V_C^{\text{am}} \geq S - Ke^{-r(T-t)}$, we have

$$V_C^{\text{am}} > S - K \quad \text{for } t < T \quad \text{and } r > 0,$$

i.e., for $\delta = 0$ American and European calls are identical: $V_C^{\text{am}} = V_C^{\text{eur}}$. In this case, the corresponding boundary conditions read as follows.

$$V_C^{\text{am}}(S_f(t), t) = S_f(t) - K, \quad (7.0.6)$$

$$\frac{\partial}{\partial S} V_C^{\text{am}}(S_f(t), t) = -1. \quad (7.0.7)$$

Black-Scholes Inequality

In the derivation of (7.0.5) early exercise was excluded. Following the lines of argumentation ‘= 0’ is now replaced by ‘ ≤ 0 ’ and we obtain *Black-Scholes-Inequality*, which holds for all (S, t) .

The inequality can be reformulated taking into account that the contact boundary S_f divides the half strip into two disjoint regulars.

$$\begin{aligned} \text{Put: } V_P^{\text{am}} &= K - S && \text{for } S \leq S_f \text{ (stop),} \\ V_P^{\text{am}} &\text{ solves (7.0.5)} && \text{for } S > S_f \text{ (hold).} \end{aligned} \quad (7.0.8)$$

$$\begin{aligned} \text{Call: } V_C^{\text{am}} &= S - K && \text{for } S \geq S_f \text{ (stop),} \\ V_C^{\text{am}} &\text{ solves (7.0.5)} && \text{for } S < S_f \text{ (hold).} \end{aligned} \quad (7.0.9)$$

This shows that also here the Black-Scholes equations have to be solved but with the additional problem of the *free boundary*.

7.1 The Binomial Method

We have already seen the binomial method for European options. Now, we give an easy and straightforward modification to American Options. It just amounts to including one projection step.

In fact, the only difference is the computation of V_{ji} , the approximation of $V(t_i, S_{ji})$ where $S_{ji} = S_0 u^j d^{i-j}$ so that (t_i, S_{ji}) can be seen as grid points. Then, we have

Theorem 7.1.1 Input: r, σ, S_0, K, M , choice of put or call

- $\Delta t = \frac{T}{M}$ u, d, p like in the European case
- $S_{00} = S_0$
- $S_{jM} = S_{00} u^j d^{M-j}$, $j = 0, 1, \dots, M$
 $S_{ji} = S_{00} u^j d^{i-j}$, $i = 1, \dots, M-1$ $j = 0, 1, \dots, y$

- $V_{j,M} = \begin{cases} (S_{jM} - K)^+ & \text{for Call} \\ (K - S_{jM})^+ & \text{for Put} \end{cases}$
- $V_{j,i} = \begin{cases} \max\{(S_{ji} - K)^+, e^{-r\Delta t}(pV_{j+1,i+1} + (1-p)V_{j,i+1})\}, & \text{Call} \\ \max\{(K - S_{ji})^+, e^{-r\Delta t}(pV_{j+1,i+1} + (1-p)V_{j,i+1})\}, & \text{Put} \end{cases}$
 $i < M$

Output: V_{00} is the approximation of $V(S_0, 0)$.

Example: See [14], Exercise 1.6.

7.2 Obstacle Problem

Another numerical method for solving the pricing problem for American Options uses the Black-Scholes inequality (see above). This can be seen as a particular instant of an *obstacle problem*, which is also common in several areas of application.

Example 7.2.1 *Given a membrane on which a force $-f$ acts on some domain $\Omega \subset \mathbb{R}^2$. The membrane is fixed on the boundary $\partial\Omega = \Gamma$ and the displacement of the membrane is bounded in Ω by a given function g (the obstacle).*

Then u is given as the solution of the following system

$$\left. \begin{aligned} -\Delta u &\geq f \\ u &\geq g \\ (-\Delta u - f)(u - g) &= 0 \end{aligned} \right\} \text{ in } \Omega, \quad u/\Gamma = 0 \quad (7.2.1)$$

Let us now subdivide Ω into the (unknown) contact zone

$$D_2 := \{x \in \Omega : u(x) = g(x)\}$$

and $D_1 := \Omega \setminus D_2$. Then we have

$$\left\{ \begin{aligned} &\bullet \text{ in } D_2 : u = g \ (\Rightarrow \Delta u = \Delta g < f) \\ &\bullet \text{ in } D_1 : u > g \ \Rightarrow -\Delta u = f, \end{aligned} \right. \quad (7.2.2)$$

i.e., the same behavior as for the Black-Scholes inequality:

if $V^{am} > \text{payoff} \Rightarrow$ Black-Scholes-equation (7.0.1,...,7.0.5) holds,

if $V^{am} = \text{payoff} \Rightarrow$ Black-Scholes-inequality.

As opposed to (7.2.2), the fomulation (7.2.1) does **not** involve the unknown D_2 , (7.2.1) is also called *linear complementary problem*. We use this for the design of a numerical method. If a solution u of (7.2.1) is determined, we can compute D_2 from it.

Variational Inequalities

It is known from the theory of partial differential equations that the classical (strong) formulation of boundary value problems is often not appropriate. The same also holds for inequalities.

Example 7.2.2 Consider the 1d elliptic PDE:

$$-u'' = f(x), x \in (0, 1), \quad u(0) = u(1) = 0,$$

which leads to the variational formulation of finding $u \in H_0^1(0, 1)$ such that

$$(\nabla u, \nabla v)_0 =: a(u, v) = (f, v)_0 \quad \forall v \in H_0^1(0, 1),$$

which is equivalent to the minimization problem

$$J(v) := \frac{1}{2}a(v, v) - (f, v)_0 \rightarrow \min \quad \text{for } v \in H_0^1(0, 1).$$

If the above minimization has to be constrained, i.e., $v \in H_0^1(0, 1)$ is replaced by a (convex) subset $K \subset H_0^1(0, 1)$, then we obtain a variational inequality (its analysis leads to the field of convex analysis).

The general form (which is also appropriate for American option pricing problems) reads as follows: Let H be a real Hilbert space and $K \subset H$ convex, $K \neq \emptyset$. Further, let $L : K \subset H \rightarrow H'$ be given. Then, one has to determine $u \in K$ such that

$$\langle L(u), v - u \rangle \geq 0 \quad \forall v \in K \tag{7.2.3}$$

where $\langle \cdot, \cdot \rangle$ is the duality pairing of H and its dual H' , i.e., for $w \in H'$, $v \in H$, the duality pairing is defined by $\langle w, v \rangle := w(v) \in \mathbb{R}$.

Remark 7.2.3 If K is a linear subspace of H , i.e., $v := u \pm z \in K$ for all $u, z \in K$. Inserting this in (7.2.3) yields on one hand $\langle L(u), z \rangle \geq 0$ and on the other hand, due to the linearity of $\langle L(u), \cdot \rangle$ that $\langle L(u), -z \rangle = -\langle L(u), z \rangle \geq 0$, i.e., we obtain $\langle L(u), z \rangle = 0$ for all $z \in K$, i.e., a variational equality.

Example 7.2.4 In Example 7.2.1, we obtain $K = \{v \in H_0^1(\Omega) : v \geq g \text{ in } \Omega \text{ a.e.}\}$ and

$$-\int_{\Omega} \nabla u \nabla (v - u) \, dx = a(u, v - u) \geq (f, v - u)_0, \quad \forall v \in K. \quad (7.2.4)$$

i.e., $L : K \rightarrow V'$ is defined on all $V > K$, i.e.,

$$\langle Lu, v \rangle = a(u, v) - \langle f, v \rangle.$$

The relation to the complementarity problem (7.2.1) is as follows. If (in addition to the above assumptions) $u \in H^2(\Omega)$, we have by integration by parts

$$\int_{\Omega} \nabla u \nabla (v - u) \, dx = \int_{\Omega} (-\Delta u)(v - u) \, dx,$$

i.e., (7.2.4) reads

$$(-\Delta u, v - u)_0 \geq (f, v - u)_0 \quad \forall v \in K \quad (7.2.5)$$

and since $u \in K$ we have $u \geq g$ in Ω a.e. Now let $\varphi \in H_0^1(\Omega)$, $\varphi \geq 0$ in Ω a.e. Then, setting $v := \varphi + u \in K$, we obtain

$$(-\Delta u, \varphi)_0 \geq (f, \varphi)_0 \quad \forall \varphi \geq 0, \quad (7.2.6)$$

i.e., $-\Delta u \geq f$ in Ω . Choosing $v := g$ in (7.2.5), we have

$$\begin{aligned} (-\Delta u - f, g - u)_0 &\geq 0 \text{ as well as} \\ \underbrace{(-\Delta u - f, g - u)_0}_{\geq 0} &\leq 0, \text{ so that in total} \end{aligned}$$

$(-\Delta u - f)(g - u) = 0$. Hence, u solves the complementary problem (7.2.1). On the other hand, let $u \in H^2(\Omega)$ solve (7.2.1), then we have from the first and second equation that

$$(-\Delta u - f, v - g)_0 \geq 0 \quad \forall v \in K$$

as well as

$$(-\Delta u - f, u - g)_0 = 0,$$

from the third equation. Next, by subtracting the latter two equations yields

$$\begin{aligned} 0 &\leq (-\Delta u - f, v - g)_0 - (-\Delta u - f, u - g)_0 \\ &= (-\Delta u - f, v - u)_0 \\ &= a(u, v - u) - (f, v - u)_0, \end{aligned}$$

i.e., (7.2.4).

Remark 7.2.5 *The equivalent minimization problem reads*

$$J(v) := \frac{1}{2}a(v, v) + (f, v)_0 \rightarrow \min_{v \in K}!$$

Hence, an obstacle problem may equivalently be formulated as

- free boundary value problem,
- linear complementary problem,
- variational inequality (if $u \in H^2(\Omega)$),
- minimization problem (if $u \in H^2(\Omega)$).

7.3 Finite Difference Methods

For notational convenience, we restrict ourselves to the 1D case and consider the complementary problem analogous to (7.2.1)

$$\left\{ \begin{array}{l} -u''(x) \geq f(x) \\ u(x) \geq g(x) \\ (-u''(x) - f(x))(u(x) - g(x)) = 0 \\ u(-1) = u(1) = 0, \quad u \in C^1[-1, 1] \end{array} \right\} \quad \forall x \in (-1, 1) \quad (7.3.1)$$

Again we introduce a simple equidistant grid

$$\Delta := \{-1 = x_0 < x_1 < \cdots < x_m = 1\}, \quad h := \frac{2}{m}, \quad m \in \mathbb{N} \\ x_i = -1 + ih, \quad 0 \leq i \leq m,$$

and use the central difference approximation

$$\begin{aligned} -u''(x_i) &\approx \frac{1}{h^2}(-u(x_{i-1}) + 2u(x_i) - u(x_{i+1})) , & f_i &:= f(x_i) \\ & & g_i &:= g(x_i) , \end{aligned}$$

i.e., we compute an approximation $u_i \approx u(x_i)$ by

$$\left\{ \begin{array}{l} u_0 = u_m = 0 \\ (-u_{i-1} + 2u_i - u_{i+1} - h^2 f_i)(u_i - g_i) = 0 \\ u_i \geq g_i \\ -u_{i-1} + 2u_i - u_{i+1} \geq h^2 f_i \end{array} \right\} \quad 1 \leq i \leq m-1, \quad (7.3.2)$$

or, in matrix-vector notation

$$\left\{ \begin{array}{l} (u - g)^T (Au - f) = 0, \\ u \geq g, \\ Au \geq f, \end{array} \right. \quad (7.3.3)$$

where $g := (g_1, \dots, g_{m-1})^T$, $f := h^2(f_1, \dots, f_{m-1})^T$, $u := (u_1, \dots, u_{m-1})^T$ and the system matrix

$$A := \begin{bmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{(m-1) \times (m-1)}.$$

Note that

$$\begin{aligned} (u - g)^T (Au - f) &= \sum_{i=1}^{m-1} (-u_{i-1} + 2u_i + u_{i+1} - h^2 f_i)(u_i - g_i) = 0 \\ \iff & (-u_{i-1} + 2u_i + u_{i+1} - h^2 f_i)(u_i - g_i) = 0 \quad \forall i \end{aligned}$$

in the case $-u_{i-1} + 2u_i + u_{i+1} \geq h^2 f_i$ and $u_i \geq g_i$ (i.e., if all signs are equal). Let us now describe a numerical (iterative) method for the solution of (7.3.3). First we note that (7.3.3) is equivalent to

$$\min\{Au - f, g - u\} = 0 \quad (\text{componentwise}). \quad (7.3.4)$$

This means either $u_i = g_u$ or $(Au)_i = f_i$, $1 \leq i \leq m-1$. We now consider the decomposition

$$A = D - L - U$$

where L is the lower left and U the upper right part of A . From now on, we assume that A is s.p.d, then

$$D_{ii} = a_{ii} > 0 \quad \forall 1 \leq i \leq m - 1 \quad (7.3.5)$$

and since

$$Au - f = D(u - D^{-1}(Lu + Uu + f))$$

the equations (7.3.4) is equivalent to

$$\min\{u - D^{-1}(Lu + Uu + f), g - u\} = 0, \quad (7.3.6)$$

or

$$u = \max\{D^{-1}(Lu + Uu + f), g\}. \quad (7.3.7)$$

The general idea is to modify appropriate iterative methods for $Au = f$ in such a way that (7.3.7) is incorporated in the iteration.

7.3.1 Classical Iterative Methods

The idea is to construct a suitable fixpoint iteration that converges towards the solution x^* of a linear system $Ax = b$.

If $A \in \mathbb{R}^{n \times n}$, choose a regular matrix $Q \in \mathbb{R}^{n \times n}$. Then

$$\begin{aligned} Ax = b &\iff Q^{-1}(Ax - b) = 0 \\ &\iff \phi(x) := \underbrace{(I - Q^{-1}A)}_{=:G} x + \underbrace{Q^{-1}b}_{=:c} = x, \end{aligned}$$

i.e., the solution of the linear system of equations is equivalent to the fixpoint problem $\phi(x) = x$. Then, Banach fixpoint theorem yields a corresponding iteration:

$$x_{k+1} = \phi(x_k) = Gx_k + c.$$

Then, we have the following standard result.

Lemma 7.3.1 *The fixpoint iteration converges for any initial value $x_0 \in \mathbb{R}^n$ if $\rho(G) < 1$, where*

$$\rho(G) := \max_{1 \leq j \leq n} |\lambda_j(G)|$$

denotes the spectral radius of G .

Proof: Consider the singular value decomposition of G

$$G = U\Sigma V^T$$

with orthogonal matrices U, V and $\Sigma = \text{diag}(\sigma_i)$, $\sigma_i = \lambda_i(G)^2 < 1$,

$$\Rightarrow \lim_{k \rightarrow \infty} \Sigma^k = 0$$

thus

$$\lim_{k \rightarrow \infty} G^k = u \left(\lim_{k \rightarrow \infty} \Sigma^k \right) V^T = 0 ,$$

which proves the claim. □

We still have the choice of the matrix Q . We describe some standard and well-known examples.

Richardson method: This corresponds to the choice $Q = \alpha I$, $\alpha \in \mathbb{R}$, i.e.,

$$x_{k+1} = x_k + \alpha(b - Ax_k)$$

The particular choice $\alpha \neq 1$ is known as *damped Richardson iteration*.

Jacobi method: With the decomposition $A = L + D + U$, where $D = \text{diag}(A)$ is the diagonal of A , one uses $Q = D^{-1}$. This results in the iteration:

$$x_{k+1} := (I - D^{-1}A)x_k + D^{-1}b = -D^{-1}(L + U)x_k + D^{-1}b.$$

For this method, we have the following convergence result.

Theorem 7.3.2 *If A is strictly diagonal dominant, i.e.,*

$$|a_{ii}| > \sum_{i \neq j} |a_{ij}| \quad \forall 1 \leq i \leq n ,$$

then the Jacobi iteration converges towards $x = A^{-1}b$ for all $x_0 \in \mathbb{R}^n$.

Proof: Follows by Lemma 7.3.1 since $G = I - D^{-1}A = -D^{-1}(L + R)$ and

$$\rho(D^{-1}(L + R)) \leq \|D^{-1}(L + R)\|_{\infty} = \max_i \sum_{i \neq j} \left| \frac{a_{ij}}{a_{ii}} \right| < 1 ,$$

which proves the theorem. □

Let us now consider the number of operations:

- $\mathcal{O}(N^2)$ for dense matrices $A \in \mathbb{R}^{N \times N}$,
- $\mathcal{O}(N)$ for sparse matrices

per step.

Gauß-Seidel method: Again, we use the decomposition $A = L + D + U$ and set $Q = D + L$, which yields the iteration:

$$\begin{aligned} x_{k+1} &= (I - (D + L)^{-1}A)x_k + (D + L)^{-1}b \\ &= -(D + L)^{-1}Ux_k + (D + L)^{-1}b, \end{aligned}$$

since

$$\begin{aligned} I - (D + L)^{-1}A &= (D + L)^{-1}(D + L - A) \\ &= (D + L)^{-1}(D + L - D - L - U). \end{aligned}$$

For this method, the following result is known.

Theorem 7.3.3 *The Gauß-Seidel method converges for all s.p.d. matrices $A \in \mathbb{R}^{n \times n}$.*

For the proof, we need some preparations. For a s.p.d. matrix $A \in \mathbb{R}^{n \times n}$, we consider the following scalar product.

$$(x, y)_A := x^T A y = (x, Ay), \quad x, y \in \mathbb{R}^n.$$

Note that $B^* = A^{-1}B^T A$ is the *A-adjoint matrix* i.e.,

$$(Bx, y)_A = (x, B^*y)_A$$

for all $y, x \in \mathbb{R}^n$. In fact, we have

$$(Bx, y)_A = x^T B^T A y = x^T A \underbrace{A^{-1}B^T A}_{=B^*} y = (x, B^*y)_A.$$

Any *A-selfadjoint* matrix B (which means that $B^* = B$) is called *A-positiv* if

$$(Bx, x)_A > 0 \quad \forall x \neq 0.$$

Lemma 7.3.4 *If $B := I - G^*G$ with $G \in \mathbb{R}^{n \times n}$, is A -positiv, then $\rho(G) < 1$.*

Proof: By assumption, B is A -positiv, i.e.

$$0 < (Bx, x)_A = (x, x)_A - (G^*Gx, x)_A = (x, x)_A - (Gx, Gx)_A$$

which means that $(x, x)_A > (Gx, Gx)_A$. Thus, we have for the A -norm

$$\|x\|_A := \sqrt{(x, x)_A}$$

that $\|x\|_A > \|Gx\|_A$. Finally

$$\rho(G) \leq \|G\|_A := \sup_{\|x\|_A=1} \frac{\|Gx\|_A}{\|x\|_A} < 1$$

since the supremum is attained due to the compactness of the unit ball with respect to $\|\cdot\|_A$, i.e. $\partial B_{1,A}(0) := \{x \in \mathbb{R}^n : \|x\|_A = 1\}$. \square

Proof of Theorem 7.3.3: Show that $B := I - G^*G$ is A -positive for $G = I - (D + L)^{-1}A$. Since $U^T = L$, we have

$$\begin{aligned} G^* &= I - A^{-1}A^T(D + L)^{-T}A \\ &= I - (D^T + L^T)^{-1}A = I - (D + U)^{-1}A, \end{aligned}$$

and thus by standard calculations

$$\begin{aligned} B &= I - G^*G = I - (I - (D + U)^{-1}A)(I - (D + L)^{-1}A) \\ &= I - I + (D + U)^{-1}A + (D + L)^{-1}A - \underbrace{(D + U)^{-1}A}_{\substack{=(D+U)^{-1}(D+U+L) \\ =I+(D+U)^{-1}L}} (D + L)^{-1}A \\ &= (D + U)^{-1}A - (D + U)^{-1}L(D + L)^{-1}A \\ &= (D + U)^{-1}D \underbrace{(D^{-1}(D + L) - D^{-1}L)}_{=I+D^{-1}L-D^{-1}L} (D + L)^{-1}A \\ &= (D + U)^{-1}D(D + L)^{-1}A. \end{aligned}$$

Hence, we obtain

$$\begin{aligned} (Bx, x)_A &= ((D + U)^{-1}D(D + L)^{-1}Ax, Ax) \\ &= (D(D + L)^{-1}Ax, \underbrace{(D + U)^{-T}Ax}_{(D+L)^{-1}}) \\ &= (D^{1/2}(D + L)^{-1}Ax, (D^{1/2}(D + L)^{-1}Ax)) \\ &= \underbrace{\| (D^{1/2}(D + L)^{-1}Ax) \|}_{\text{regular}} > 0 \end{aligned}$$

for $x \neq 0$ which shows that B is s.p.d. Thus, the claim follows from Lemma 7.3.4. \square

Relaxation Methods

For Gauß-Seidel, we have $G = I - (D + L)^{-1}A$. In order to speed up the convergence, one can introduce an additional parameter $\omega > 0$ and obtain the iteration matrix

$$G_\omega := I - \left(\frac{1}{\omega}D + L \right)^{-1} A ,$$

which means in particular, that for $\omega = 1$, we obtain the above mentioned Gauß-Seidel method. Thus, we obtain the iteration

$$x^{(k+1)} = \left(I - \left(\frac{1}{\omega}D + L \right)^{-1} A \right) x_k + \left(\frac{1}{\omega}D + L \right)^{-1} b .$$

The method is applied in practice as follows for a given iterate $x^{(k)}$

$$\left\{ \begin{array}{l} \text{For } i = 1, \dots, N \\ z_i^{(k+1)} = \frac{1}{a_{ii}} \left[\underbrace{- \sum_{m < i} a_{im} x_m^{(k+1)} - \sum_{m > i} a_{im} x_m^{(k)} + b_i}_{= (-Lx^{(k+1)} - Ux^{(k)} + b)_i} \right] \\ x_i^{(k+1)} = x_i^{(k)} + \omega (z_i^{(k+1)} - x_i^{(k)}) . \end{array} \right. \quad (7.3.8)$$

This can be seen as follows:

$$\begin{aligned} a_{ii}x_i^{(k+1)} &= a_{ii}x_i^{(k)} + \omega \left[- \sum_{m < i} a_{im}x_m^{(k+1)} - \sum_{m > i} a_{im}x_m^{(k)} + b_i - a_{ii}x_i^{(k)} \right] \\ \Leftrightarrow Dx^{(k+1)} &= Dx^{(k)} + \omega(-Lx^{(k+1)} - Ux^{(k)} + b - Dx^{(k)}) \\ \Leftrightarrow (D + \omega L)x^{(k+1)} &= (D - \omega \overbrace{(U + D)}^{=A-L})x^{(k)} + \omega b \\ \Leftrightarrow \omega \left(\frac{1}{\omega}D + L \right) x^{(k+1)} &= \omega \left[\left(\frac{1}{\omega}D + L \right) - \omega A \right] x^{(k)} + \omega b \end{aligned}$$

\Leftrightarrow

$$x^{(k+1)} = \left(I - \left(\frac{1}{\omega} D + L \right)^{-1} A \right) x^{(k)} + \left(\frac{1}{\omega} D + L \right)^{-1} b.$$

For $\omega < 1$, this is called a *damped iteration*, for $1 < \omega < 2$ it is called *over-relaxed*, the method is also known as SOR (*successive over relaxation*). Details can be found e.g. in [15] 2, §8.

Theorem 7.3.5 (Ostrowski, Reich) *For any s.p.d. matrix $A \in \mathbb{R}^{n \times n}$, the SOR method converges for all $0 < \omega < 2$.*

The proof can be found in any standard textbook on Numerical Analysis. The question naturally arises what might be an optimal choice for the parameter ω .

Definition 7.3.6 *A matrix $A \in \mathbb{R}^{n \times n}$ is called consistently ordered if the eigenvalues of the matrices*

$$J(\alpha) := D^{-1}(\alpha L + \alpha^{-1}U) \quad (\alpha \neq 0) \tag{7.3.9}$$

are independent of α , if $A = L + D + U$.

The following theorem can be found e.g. in [15].

Theorem 7.3.7 *If A is consistently ordered, then*

$$\rho(G_1) = \rho(J)^2, \quad J = J(1),$$

where $G_1 = I - (D + L)^{-1}A = -(D + L)^{-1}U$ is the iteration matrix of the Gauß-Seidel method. \square

Note that $-J(1) = -D^{-1}(L + U)$ is the iteration method of Jacobi, thus Theorem 7.3.7 says that the Jacobi method roughly needs the double number of iterations than Gauß-Seidel (if A is consistently ordered).

Now the optimal parameter is characterized by

$$\rho(G_{\omega_{\text{opt}}}) = \min_{\omega \in \mathbb{R}} \rho(G_\omega) = \min_{0 < \omega < 2} \rho(G_\omega)$$

and the following result is known (see again [15]).

Theorem 7.3.8 (Young, Varga) *Let A be consistently ordered and assume that $J = J(1)$ has only real eigenvalues such that $\rho(J) < 1$ (see Lemma 7.3.1). Then*

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho(J)^2}} , \quad \rho(G_{\omega_{opt}}) = \omega_{opt} - 1 .$$

□

Remark 7.3.9 *Note that tridiagonal- and block-tridiagonal-matrices are consistently ordered which can easily be verified.*

7.3.2 Projected SOR-method for Complementary Problems

Now, we modify the above described SOR-method for solving the complementary problem

$$(Au - f)^T(u - g) = 0 , \quad u \geq g , \quad Au \geq f \quad (7.3.10)$$

which we have seen to be equivalent to (7.3.7), i.e.,

$$u = \max\{D^{-1}(Lu + Uu + f), g\} ,$$

if the matrix $A = D - L - U$ is s.p.d. We add a projection step in the SOR-method(7.3.8) (note: there we have used the decomposition $A = D + L + U$):

$$\left\{ \begin{array}{l} \text{For } i = 1, \dots, N \text{ do} \\ z_i^{(k+1)} = \frac{1}{a_{ii}}(Lu^{(k+1)} + Uu^{(k)} + f)_i \\ u_i^{(k+1)} = \max\{u_i^{(k)} + \omega(z_i^{(k+1)} - u_i^{(k)}), g_i\} \end{array} \right. \quad (7.3.11)$$

which is called *projected SOR-method*.

We aim to prove that (7.3.11) converges towards the solution u of (7.3.10). We need some preparations. The following proofs are taken from [5].

Lemma 7.3.10 *The problem (7.3.10) is equivalent to*

$$u \geq g , \quad J(u) = \min_{v \geq g} J(v) \quad (7.3.12)$$

where $J(v) := \frac{1}{2}v^T Av = f^T v$, if A is s.p.d.