Chapter 2
Simulation and Continuous Optimization

Oliver Kolb and Jens Lang

Abstract In this chapter we consider the solution of the model equations of water supply networks and continuous optimal control tasks. We begin with the description of our simulation tool in Sect. 2.1, in particular the numerical treatment of the water hammer equations. This includes the description of the implemented discretization scheme together with a stability and convergence analysis. As we will see, the applied scheme perfectly matches with the properties of the water hammer equations and thus builds a useful foundation for the solution of the entire model equations as well as optimal control tasks.

In Sect. 2.2 we consider the computation of sensitivity information, which is necessary for the application of gradient-based optimization techniques. Here, we follow a first-discretize approach to derive adjoint equations. Due to the special structure of the considered problems, very efficient algorithms can be applied.

Finally, Sect. 2.3 deals with the problem of singularities in the model equations of water supply networks. Here, a physically motivated regularization approach is applied and also extended to be applicable in an adjoint calculus.

2.1 Numerical Solution of the Model Equations

In this section, we describe our simulation tool, which numerically solves the underlying model equations. The main structure of this tool is described in Sect. 2.1.1. Here, we assume that the discretization of the model equations in time and space is given.

The water hammer equations are an integral part of the entire model of water supply networks. As we will see, this system of partial differential equations is hyperbolic. The numerical solution of hyperbolic PDEs demands great care regarding the discretization scheme, which crucially depends on the properties of the underlying equations. After recapitulating the basic properties of the water hammer equations in Sect. 2.1.2, we will describe the applied discretization scheme in Sect. 2.1.3 and give stability and convergence results.

2.1.1 Network Equations

The first step towards the solution of the model equations is an appropriate discretization. The treatment of the water hammer equations is described in detail in

A. Martin et al. (eds.), *Mathematical Optimization of Water Networks*,
International Series of Numerical Mathematics 162,
DOI 10.1007/978-3-0348-0436-3_2,

Sect. 2.1.3. The same time discretization is applied to the other components, modelled by algebraic and ordinary differential equations. The latter are discretized with one-step methods.

Let $t_0 < t_1 < \cdots < t_N$ be the time steps of the discretization. The application of the discretization schemes to the model equations yields a coupled system of (nonlinear) algebraic equations $E(y, u)$, which depends on state variables

$$y^T = \left(y(t_0)^T, y(t_1)^T, \ldots, y(t_N)^T\right),$$

like pressure head and flow rates, and control variables

$$u^T = \left(u(t_0)^T, u(t_1)^T, \ldots, u(t_N)^T\right),$$

e.g. the speed of pumps. Boundary and coupling conditions are already included in $E(y, u)$. Now, the simulation task consists of solving these equations for a given initial state $y(t_0)$ and control variables for all time steps. Due to the time-dependent structure, this set of equations can be partitioned and solved for $y(t_j)$ time step by time step ($j = 1, \ldots, N$), resulting in subsets of E of the form

$$F\left(t_{j-1}, t_j, y(t_{j-1}), y(t_j), u(t_{j-1}), u(t_j)\right) = 0.$$

While explicit dependencies may be solved in advance, the remaining (implicit) equations have to be solved with Newton's method. Here, we can exploit the sparsity structure of the underlying Jacobian matrix by using an appropriate solver for the sets of linear equations [7]. Unfortunately, the discretized model equations of water supply networks do not always yield unique solutions. The treatment of the underlying singularities is described in Sect. 2.3.

2.1.2 Properties of the Water Hammer Equations

The water hammer equations play an integral role in the modelling of water supply networks. The purpose of this section is to collect some properties of particular interest.

The water hammer equations are given by

$$\begin{aligned} &\frac{\partial}{\partial t} h + \frac{a^2}{gA} \frac{\partial}{\partial x} Q = 0, \\ &\frac{\partial}{\partial t} Q + gA \frac{\partial}{\partial x} h = -\lambda(Q) \frac{Q|Q|}{2dA} \end{aligned} \tag{2.1}$$

and can be written in the general form of a balance law

$$\frac{\partial}{\partial t} w + \frac{\partial}{\partial x} f(w) = g(w)$$

with $w = \binom{h}{Q}$ and

$$f(w) = \begin{pmatrix} \frac{a^2}{gA} Q \\ gAh \end{pmatrix}, \qquad g(w) = \begin{pmatrix} 0 \\ -\lambda(Q) \frac{Q|Q|}{2dA} \end{pmatrix}.$$

Obviously, $f(w)$ is linear in w and we have

$$\frac{\partial}{\partial x} f(w) = \underbrace{\begin{pmatrix} 0 & \frac{a^2}{gA} \\ gA & 0 \end{pmatrix}}_{= \frac{\partial}{\partial w} f(w)} \frac{\partial}{\partial x} w.$$

A short calculation yields

$$\lambda_{1,2} = \pm a$$

for the eigenvalues of $\frac{\partial}{\partial w} f(w)$. Thus, the water hammer equations form a hyperbolic system of PDEs with constant characteristic speeds.

Another important property of the water hammer equations is the dissipativity of the source term $g(w)$: The eigenvalues of

$$\frac{\partial}{\partial w} g(w) = \begin{pmatrix} 0 & 0 \\ 0 & -\frac{|Q|}{2dA} (\lambda'(Q) Q + 2\lambda(Q)) \end{pmatrix}$$

are

$$\mu_1 = -\frac{|Q|}{2dA} \big(\lambda'(Q) Q + 2\lambda(Q)\big) < 0, \qquad \mu_2 = 0.$$

In practical computations, often the stationary limit of the water hammer equations is used. Setting the time derivatives to zero yields that the discharge is constant (in space),

$$Q(x) \equiv Q, \tag{2.2}$$

and a linearly decreasing pressure head in flow direction,

$$h(x_1) - h(x_0) = -\lambda(Q) \frac{Q|Q|}{2gdA^2} (x_1 - x_0). \tag{2.3}$$

2.1.3 Implicit Box Scheme

For the discretization of the water hammer equations, we apply an implicit box scheme. The main drawback of explicit methods in the context of hyperbolic PDEs is the stepsize restriction due to the CFL condition, which is of the form

$$\Delta t \leq \alpha \frac{\Delta x}{\lambda_{max}} \tag{2.4}$$

with some positive $\alpha \in \mathbb{R}$. Here, λ_{max} denotes the spectral radius of the Jacobian matrix of the flux function. Regarding the water hammer equations, the CFL condition is very restrictive since λ_{max} equals the speed of sound (in water). Thus, very fine time discretizations would be necessary for stability reasons, while an appropriate resolution of the typically moderate dynamics in the daily operation of water supply networks would allow much larger time steps.

We now formulate the applied scheme for balance laws of the form

$$\frac{\partial}{\partial t} w + \frac{\partial}{\partial x} f(w) = g(w), \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+ \tag{2.5}$$

with given initial data

$$w(x,0) = w_0(x), \quad x \in \mathbb{R}. \tag{2.6}$$

To approximate (weak) solutions of (2.5)–(2.6), we choose a spatial mesh size Δx, a time grid size Δt and introduce a piecewise constant function $\tilde{w}(x,t)$ defined by

$$\tilde{w}(x,t) = w_j^n \quad \text{for } (x,t) \in I_j \times J_n \tag{2.7}$$

with $I_j = [(j-0.5)\Delta x, (j+0.5)\Delta x)$ and $J_n = [n\Delta t, (n+1)\Delta t)$. For the computation of the approximate values $w_j^n \approx w(j\Delta x, n\Delta t)$, we consider the implicit box scheme

$$\begin{aligned} \frac{w_{j-1}^{n+1} + w_j^{n+1}}{2} = \frac{w_{j-1}^n + w_j^n}{2} - \frac{\Delta t}{\Delta x}\left(f\left(w_j^{n+1}\right) - f\left(w_{j-1}^{n+1}\right)\right) \\ + \Delta t \frac{g(w_{j-1}^{n+1}) + g(w_j^{n+1})}{2}. \end{aligned} \tag{2.8}$$

As initial conditions, we set

$$w_j^0 = \int_{I_j} w_0(x)\, dx. \tag{2.9}$$

When implementing this method for a scalar balance law on a finite grid $x_l < x_{l+1} < \cdots < x_{r-1} < x_r$, we get $r-l$ equations for $r-l+1$ variables. So, we have to impose boundary conditions at exactly one boundary, depending on the characteristic direction, i.e., on the sign of f'. In order that the proposed scheme may work, we have to assume that the sign of f' does not change over the computational domain. The generalization for systems of balance laws is that the signature of the characteristic directions does not change. This assumption is often satisfied for subsonic flows and also holds for the water hammer equations (2.1).

We mention that for $g \equiv 0$ and $w^n, w^{n+1} \in L^1(\mathbb{Z})$, the scheme (2.8) is conservative. Moreover, it can be easily shown that the proposed scheme is exact in the stationary case (2.2)–(2.3) of the water hammer equations. Next, we give some further results for the applied scheme in the scalar case, which have already been published in [12], where also the proofs can be found.

First, it can be shown that the box scheme admits a unique solution in $L^1(\mathbb{Z})$ in every time step. For this, we assume $f' \geq \lambda_{min} > 0$. Analogously, Proposition 1 and the following propositions hold in the case $f' \leq -\lambda_{min} < 0$.

Proposition 1 (Existence and Uniqueness) *For* $w^n \in L^1(\mathbb{Z})$, $f, g \in C^1(\mathbb{R})$, $g(0) = 0$, $g' \leq 0$, $f' \geq \lambda_{min} > 0$ *and* $\frac{\Delta t}{\Delta x} \geq \frac{1}{2\lambda_{min}}$, *scheme* (2.8) *admits a unique solution* $w^{n+1} \in L^1(\mathbb{Z})$.

In Proposition 1, we have introduced the requirement $\Delta t \geq \Delta x/(2\lambda_{min})$. In contrast to the CFL condition (2.4), which determines an upper bound for the time grid size Δt, the implicit structure of the scheme leads to a lower bound for the time grid size.

Motivated by the well-known results of Kružkov [13], the following stability results can be shown:

Proposition 2 (Stability) *Let* $w^n, v^n \in L^\infty(\mathbb{Z}) \cap L^1(\mathbb{Z}) = L^1(\mathbb{Z})$, $f, g \in C^1(\mathbb{R})$, $g(0) = 0$ *and* $g' \leq 0$. *Then, scheme* (2.8) *has the following stability properties*:

(1) *If* $\frac{\Delta x}{\Delta t} \leq 2f' + \Delta x g'$, *then* $\|w^{n+1}\|_{L^\infty(\mathbb{Z})} \leq \|w^n\|_{L^\infty(\mathbb{Z})}$.
(2) *If* $\frac{\Delta x}{\Delta t} \leq 2f'$, *then* $\|w^{n+1} - v^{n+1}\|_{L^1(\mathbb{Z})} \leq \|w^n - v^n\|_{L^1(\mathbb{Z})}$ *and* $TV(w^{n+1}) \leq TV(w^n)$.

The requirement in (1) can even be weakened to $\frac{\Delta x}{\Delta t} \leq 2f'$ under mild additional assumptions. Next, in analogy to the Lax-Wendroff-Theorem (see e.g. [14], pp. 239ff.), it can be shown:

Proposition 3 *Let* $(w^{(k)})_{k\in\mathbb{N}}$ *be a sequence constructed by scheme* (2.8)–(2.9) *and converging in* $L^1_{loc}(\mathbb{R} \times \mathbb{R}_+)$ *with* $\Delta t^{(k)}, \Delta x^{(k)} \overset{k\to\infty}{\longrightarrow} 0$. *Then, the limit* $\hat{w} = \lim_{k\to\infty} w^{(k)}$ *is a weak solution of the Cauchy problem* (2.5)–(2.6).

Finally, assuming the stability properties

$$\begin{aligned} \left\|w^{n+1}\right\|_{L^\infty(\mathbb{Z})} &\leq \left\|w^n\right\|_{L^\infty(\mathbb{Z})}, \\ \left\|w^{n+1}\right\|_{L^1(\mathbb{Z})} &\leq \left\|w^n\right\|_{L^1(\mathbb{Z})}, \\ TV\left(w^{n+1}\right) &\leq TV\left(w^n\right), \end{aligned} \tag{2.10}$$

which can for instance be achieved by fulfilling the requirements of Proposition 2, convergence to the so-called entropy solution can be shown:

Proposition 4 (Convergence to Entropy Solution) *Let* $w_0 \in L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$, $f, g \in C^1(\mathbb{R})$, $g(0) = 0$, $g' \leq 0$, $f' \geq \lambda_{min} > 0$ *and* $TV(u_0) < \infty$. *Let* $(w^{(k)})_{k\in\mathbb{N}}$ *be a sequence constructed by scheme* (2.8)–(2.9), *fulfilling the stability properties* (2.10) *and with* $\Delta t^{(k)}, \Delta x^{(k)} \overset{k\to\infty}{\longrightarrow} 0$, *where* $r = \frac{\Delta t^{(k)}}{\Delta x^{(k)}} \geq \frac{1}{2\lambda_{min}}$. *Then, the limit*

$\hat{w} = \lim_{k\to\infty} w^{(k)}$ *exists in* $L^1_{loc}(\mathbb{R}\times\mathbb{R}_+)$ *and is the entropy solution of the Cauchy problem* (2.5)–(2.6).

2.2 Adjoint Calculus

The last section was concerned with the solution of the simulation task. The next step is to determine the control u in such a way that a given objective function is optimized while certain constraints have to be fulfilled. Therefore, we consider the following optimal control problem:

$$\begin{aligned} \min_u\ & f\big(y(u),u\big) \\ \text{s.t. } & g\big(y(u),u\big) \geq 0 \\ & h\big(y(u),u\big) = 0 \\ & u_{min} \leq u \leq u_{max} \end{aligned} \tag{2.11}$$

with state vector y, control vector u, objective function f, inequality constraints g and equality constraints h. The state vector is assumed to be a function of the control vector, that is, for a given control u the state y is uniquely determined. As described in Sect. 2.1.1, the state vector results from solving a (nonlinear) set of equations $E(y,u)=0$ for y. For this reason, the functions f, g and h can also be considered as functions solely depending on the control u. In fact, the state variables are not visible for the optimization tools we have linked to our software, their interface only contains the control variables.

We want to solve (2.11) with gradient-based optimization methods like DONLP2 [17, 18], IPOPT [19] and KNITRO [5]. While the simulation tool enables us to evaluate the objective function and the constraints for a given control u, we still have to provide sensitivity information for all functions with respect to the control. Adjoint calculus is a very efficient way to compute the so-called *reduced gradients*. In principle, there are two different ways to compute the desired information via adjoint equations as shown in Fig. 2.1. Starting with the model equations, one may first derive (analytically) adjoint equations and apply an appropriate discretization scheme afterwards. The second possibility is to derive adjoint equations based on the discretized model equations.

Since both approaches have their advantages and disadvantages, we apply both in our software. Nevertheless, there is a strong emphasis on the second approach because the necessary components are much easier to implement. Further details of this approach are provided in the following sections.

2.2.1 The First-Discretize Approach

We consider the computation of the reduced gradient of an arbitrary scalar function $f(y(u),u)$ via adjoint equations derived by a first-discretize approach. As above

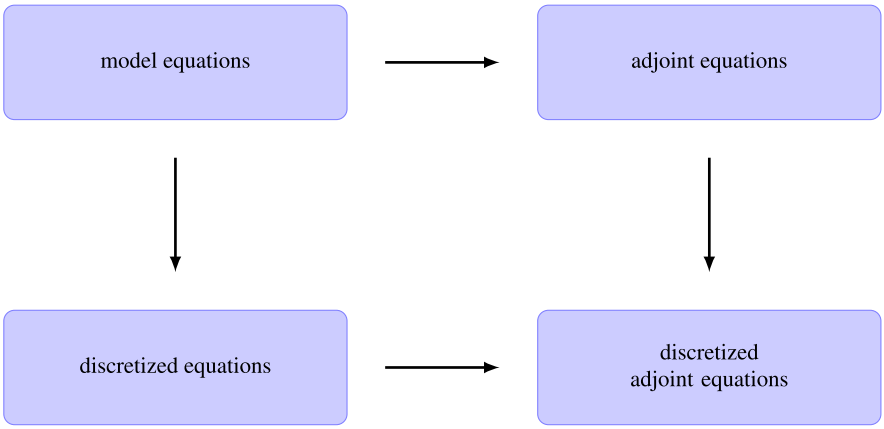

Fig. 2.1 Two ways from the model equations to the discretized adjoint equations

$y(u)$ is considered to be the unique solution of $E(y,u)=0$. Of course, the described procedure is not only valid for f being our objective function but any of the equality or inequality constraints in (2.11).

To derive the adjoint equations, which are necessary for the computation of $\frac{d}{du}f(y(u),u)$, we introduce the Lagrange function

$$L(y,u)=f(y,u)+\xi^T E(y,u), \tag{2.12}$$

where ξ is the so-called *adjoint state*. With $y=y(u)$, basic transformations of (2.12) lead to

$$\begin{aligned}\frac{d}{du}f\big(y(u),u\big)&=\frac{d}{du}L\big(y(u),u\big)-\frac{d}{du}\xi^T\underbrace{E\big(y(u),u\big)}_{=0}=\frac{d}{du}L\big(y(u),u\big)\\&=\underbrace{\frac{\partial}{\partial y}L\big(y(u),u\big)}_{\stackrel{!}{=}0\Rightarrow\xi}\frac{dy}{du}+\frac{\partial}{\partial u}L\big(y(u),u\big)=\frac{\partial}{\partial u}L\big(y(u),u\big)\\&=\frac{\partial}{\partial u}f\big(y(u),u\big)+\xi^T\frac{\partial}{\partial u}E\big(y(u),u\big).\end{aligned} \tag{2.13}$$

Thus, we have reduced the task of computing the total derivative of f with respect to u to the computation of the partial derivatives of f and E with respect to u and solving the system of *adjoint equations*:

$$\frac{\partial}{\partial y}L\big(y(u),u\big)=\frac{\partial}{\partial y}f\big(y(u),u\big)+\xi^T\frac{\partial}{\partial y}E\big(y(u),u\big)\stackrel{!}{=}0$$

$$\Leftrightarrow \quad \underbrace{\left(\frac{\partial}{\partial y}E\big(y(u),u\big)\right)^T}_{\text{independent of } f}\xi = -\left(\frac{\partial}{\partial y}f\big(y(u),u\big)\right)^T. \tag{2.14}$$

It is important to notice that (2.14) is a linear system and the matrix $\frac{\partial}{\partial y}E(y(u),u)$ is independent of the function f. Therefore, this matrix and any decomposition of it computed for solving (2.14) only needs to be computed once.

After having solved (2.14), we get our reduced gradient from (2.13):

$$\frac{d}{du}f\big(y(u),u\big) = \frac{\partial}{\partial u}f\big(y(u),u\big) + \xi^T \underbrace{\frac{\partial}{\partial u}E\big(y(u),u\big)}_{\text{independent of } f}.$$

Here, the matrix $\frac{\partial}{\partial u}E(y(u),u)$ is independent of f and therefore only needs to be computed once, independent of the number of computed gradients.

2.2.2 Application to Time-Dependent Problems

In the case of time-dependent control problems, the task (2.11) and therewith the reduced gradient (2.13) and the adjoint system (2.14) have a very special structure.

Let us begin with the state defining function $E(y,u)$. As described in Sect. 2.1.1, we start in a certain state $y(t_0) = y_0$. From any state $y(t_j)$, we come to the next state $y(t_{j+1})$ by solving a set of equations of the following form:

$$F\big(t_{old}, t_{new}, y(t_{old}), y(t_{new}), u(t_{old}), u(t_{new})\big) = 0. \tag{2.15}$$

Altogether, we have

$$E(y,u) = \begin{pmatrix} y(t_0) - y_0 \\ F(t_0, t_1, y(t_0), y(t_1), u(t_0), u(t_1)) \\ \vdots \\ F(t_{N-1}, t_N, y(t_{N-1}), y(t_N), u(t_{N-1}), u(t_N)) \end{pmatrix} = 0. \tag{2.16}$$

For the matrix in the adjoint system, we get

$$\frac{\partial}{\partial y}E(y,u) = \begin{pmatrix} I & & & & \\ A_1 & B_1 & & & \\ & A_2 & B_2 & & \\ & & \ddots & \ddots & \\ & & & A_N & B_N \end{pmatrix}$$

with

$$A_j = \frac{\partial}{\partial y_{old}}F\big(t_{j-1}, t_j, y(t_{j-1}), y(t_j), u(t_{j-1}), u(t_j)\big)$$

and

$$B_j = \frac{\partial}{\partial y_{new}} F\big(t_{j-1}, t_j, y(t_{j-1}), y(t_j), u(t_{j-1}), u(t_j)\big).$$

For an arbitrary scalar function f, the set of adjoint equations (2.14) then reads

$$\begin{pmatrix} I & A_1^T & & & \\ & B_1^T & A_2^T & & \\ & & B_2^T & \ddots & \\ & & & \ddots & A_N^T \\ & & & & B_N^T \end{pmatrix} \begin{pmatrix} \xi(t_0) \\ \xi(t_1) \\ \vdots \\ \xi(t_N) \end{pmatrix} = - \begin{pmatrix} \frac{\partial}{\partial y_0} f(y,u)^T \\ \frac{\partial}{\partial y_1} f(y,u)^T \\ \vdots \\ \frac{\partial}{\partial y_N} f(y,u)^T \end{pmatrix}. \tag{2.17}$$

Here, the partial derivatives $\frac{\partial}{\partial y_j}$ refer to the blockwise partitioning of the state vector according to the time steps.

Due to the blockwise bidiagonal structure of the matrix $\frac{\partial}{\partial y} E(y,u)$, the linear system (2.17) can be solved backwards in time and blockwise, reducing the size of the systems to be solved:

$$\begin{aligned} \xi(t_N) &= -\big(B_N^T\big)^{-1} \frac{\partial}{\partial y_N} f(y,u)^T, \\ &\vdots \\ \xi(t_j) &= -\big(B_j^T\big)^{-1} \left(\frac{\partial}{\partial y_j} f(y,u)^T + A_{j+1}^T \xi(t_{j+1}) \right), \\ &\vdots \\ \xi(t_0) &= -\left(\frac{\partial}{\partial y_0} f(y,u)^T + A_1^T \xi(t_1) \right). \end{aligned}$$

Besides the structure of the matrix $\frac{\partial}{\partial y} E(y,u)$, the right-hand side of (2.17) typically also features a special structure. Further benefit can be made out of the structure of $\frac{\partial}{\partial u} E(y,u)$. Similar to the structure of $\frac{\partial}{\partial y} E(y,u)$, we get

$$\frac{\partial}{\partial u} E(y,u) = \begin{pmatrix} 0 & \dots & \dots & \dots & 0 \\ \frac{\partial}{\partial u_{old}} E_1 & \frac{\partial}{\partial u_{new}} E_1 & & & \\ & \frac{\partial}{\partial u_{old}} E_2 & \frac{\partial}{\partial u_{new}} E_2 & & \\ & & \ddots & \ddots & \\ & & & \frac{\partial}{\partial u_{old}} E_N & \frac{\partial}{\partial u_{new}} E_N \end{pmatrix},$$

where E_j abbreviates $F(t_{j-1}, t_j, y(t_{j-1}), y(t_j), u(t_{j-1}), u(t_j))$. Since the first block of rows equals zero, there is no need to compute $\xi(t_0)$ for the evaluation of the reduced gradient via (2.13). This result is not surprising, because the initial state

$y(t_0)$ is given and therefore does not depend on the control. Moreover, we usually have $\frac{\partial}{\partial u_{old}} E_j = 0$. In this case, $\frac{\partial}{\partial u} E(y, u)$ reduces to a block-diagonal matrix.

2.3 Singularities

As already mentioned in Sect. 2.1.1, the model equations of a water supply networks may contain non-unique solutions for certain constellations of elements and control states. Naturally, non-unique solutions cause problems when trying to solve the discretized model equations. In Newton's method, we are confronted with singular or at least ill-conditioned Jacobian matrices. But it is possible to introduce a physically reasonable regularization of the underlying matrices, which turns out to be a Tychonoff-like regularization. The presented results have already been published in [11].

2.3.1 Introduction

The first algorithm to determine pressure heads and flows for a networked system in the steady state case was published in 1936 [6]. Meanwhile, a variety of software packages has been implemented, e.g. KANET [1], STANET [2] and EPANET [15]. The latter one is released as freeware by the United States Environmental Protection Agency, broadly accepted, and often a core part of proprietary packages. But EPANET and also other codes have difficulties with certain constellations of control devices. Several problem cases have been published by Simpson in 1999 [16]. Meanwhile, the EPANET software copes with all of them but many recent publications still report about new cases where it fails or computes wrong results, e.g. [4, 9].

One underlying problem can be explained by a very simple example: Consider two closed valves as shown in Fig. 2.2. The equations modelling the pressure heads h_l and h_r and the flow rates q_l and q_r at the connection of the two valves are given as follows,

$$F(h_l, q_l, h_r, q_r) = \underbrace{\begin{pmatrix} h_l - h_r \\ q_l - q_r \\ q_l \\ q_r \end{pmatrix}}_{=: b(h_l, q_l, h_r, q_r)} \stackrel{!}{=} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{2.18}$$

Here, the pressure head between the two valves is not uniquely determined by the model equations. We only claim that h_l equals h_r. From the practical point of view, we might not be interested in the "real" pressure values between the two closed valves, but in a dynamic or quasi-stationary numerical simulation, we would expect the pressure variables to keep the same or at least similar values as in the previous time step.

Of course, one could cope with the non-uniqueness in the mentioned example but the situation becomes more difficult for large networks and especially when devices

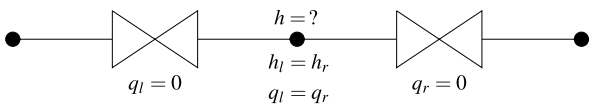

Fig. 2.2 Two closed valves—the model equations do not yield a unique solution

are state-controlled so that "truncated" parts are not known a priori. Anyway, we expect getting into trouble when solving the in general nonlinear model equations of a water supply network with Newton's method in situations like above.

The nature of the non-uniqueness in our small example is in close correlation with the Jacobian matrix of the model equations:

$$A(h_l, q_l, h_r, q_r) = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.19}$$

The Jacobian matrix is singular, and obviously, the vector $v_0 = (1, 0, 1, 0)$ is an element of the kernel of A. Thus, when solving $A\delta = b$ in Newton's method, we have to cope with the singularity of A, and moreover, an arbitrary multiple of v_0 could be added to any solution δ, which corresponds to arbitrary but equal values for h_l and h_r.

As already mentioned, from the practical point of view, we might not be interested in the pressure values between the two closed valves, but at least, we have to ensure that the underlying singularity does not impede the solution process of the whole system of model equations. Moreover, we would prefer the pressure variables to change as little as possible.

In literature, there are different approaches to handle the shown problem in general. Alvarez et al. [4] propose to add virtual tanks in the network. In [8], Deuerlein proposes a reformulation of the model equations in the form of a minimization problem. For the analysis of the resulting model, he also uses game theory. Hähnlein [10] uses basically the same modelling as we do. For solving the sets of linear equations in Newton's method, he applies singular value decomposition.

Regarding the example problem, solving $A\delta = b$ with an SVD has exactly the desired effect: We get the solution of the linear system of equations with minimal norm—the solution component in the nullspace of A (multiples of the vector v_0) equals zero. Since this property of using an SVD for solving sets of linear equations holds in a more general setting, singular value decomposition seems to be a good approach. While we get a solution of the linear system of equations (if one exists), the correction terms for the "critical variables" (here h_l and h_r) are kept small. Moreover, we may eliminate small singular values in the SVD of the matrix to stabilize the solution process. Besides all those advantages of using singular value decomposition, the price to pay is the huge computational effort.

Typically, there are various points in water supply networks where the pressure variables may not be uniquely determined and thus may be "critical" in the solution process. Without proper treatment, one gets very large correction terms δ in

Newton's method or no solution at all for the linear system of equations, and either usually leads to a failure of the method.

In the next section, we will present our approach to the solution of the introduced class of problems. In the general setting, parts of the solution of the underlying sets of linear equations are uniquely determined while there may be degrees of freedom in other parts. We want to keep those solution components as small as possible while maintaining a useful solution.

Our approach is based on the QR decomposition of a modified matrix $\tilde{A}$. Since we know the critical variables in our applications, the original matrix A is extended with additional rows in such a way that the resulting matrix has full rank. The basic idea behind this approach is to penalize corrections in critical variables. Although we increase the size of the underlying matrix, the application of a QR decomposition to the modified matrix compared to the SVD of the original matrix results in an enormous speed-up. But this is not the only contribution we make to the simulation task of water supply networks. Additionally, we use the decomposition of the modified matrix to efficiently compute sensitivity information with respect to a given target functional in Sect. 2.3.3.

2.3.2 Theoretical Analysis—Forward Direction

We consider the same setting as in Sects. 2.1.1 and 2.2: The discretization of the model equations of the whole water supply network yields a coupled system of nonlinear algebraic equations $E(y, u)$, which can be split up according to (2.16). During the solution process with Newton's method, we have to compute corrections δ by solving a linear system of equations of the form

$$A\delta = b \quad \Leftrightarrow \quad A\delta - b = 0 \tag{2.20}$$

with A being an $n \times n$ matrix. If A is singular (or ill-conditioned), we cannot make use of an LU decomposition of A in order to solve (2.20). Instead, we reformulate (2.20) as linear least squares problem

$$\min_{\delta} \|A\delta - b\|_2^2. \tag{2.21}$$

This problem can always be solved with a singular value decomposition of A. The SVD yields the (unique) solution δ^* of (2.21) where additionally $\|\delta^*\|_2$ is minimal among all solutions. In general, there is a residual $r^* = A\delta^* - b$.

Linear least squares problems can also be solved via a QR decomposition of the underlying matrix if it has full rank. Therefore, we consider the modified problem

$$\min_{\delta} \|\tilde{A}\delta - \tilde{b}\|_2^2 \tag{2.22}$$

with $\tilde{A} = \begin{pmatrix} A \\ B_s \end{pmatrix}$ and $\tilde{b} = \begin{pmatrix} b \\ 0 \end{pmatrix}$.

Here, B_s is a $k \times n$ matrix (with $k \leq n$) where in each row, there is exactly one nonzero entry $s > 0$ and at most one entry in every column, for example,

$$B_s = \begin{pmatrix} s & 0 & 0 & 0 & 0 \\ 0 & s & 0 & 0 & 0 \\ 0 & 0 & 0 & s & 0 \end{pmatrix}. \tag{2.23}$$

Let I_B be the set of column indices of the nonzero entries in B_s (in the example $I_B = \{1, 2, 4\}$). By adding additional rows to A, we can achieve that $\tilde{A}$ has full rank, and accordingly, the modified minimization problem (2.22) can be solved via a QR decomposition of $\tilde{A}$.

There are several advantages of using a singular value decomposition for solving the original problem (2.21):

1. If A is regular, $\delta^* = A^{-1}b$ and $r^* = 0$.
2. If A is not regular, $\|\delta^*\|_2$ is minimal among all solutions.
3. By eliminating small singular values, the solution process can be stabilized.

The main disadvantage of using SVD is the computational effort. Typically, the singular value decomposition is computed in two steps. First, the matrix is reduced to a bidiagonal matrix, and afterwards, the SVD of the bidiagonal matrix is computed by an iterative method up to a certain precision. In one of our real life applications, we have a 766×766 matrix with 1774 nonzero entries. For the computation of the SVD, 9.68 seconds are needed using MATLAB [3].

For the same example, the QR factorization of the corresponding modified matrix (1018×766 with 2026 nonzero entries) takes only 13 milliseconds. Hence, from the computational point of view, we prefer a QR decomposition to solve the modified problem (2.22) instead of solving (2.21) with an SVD. The results computed for the modified task (2.22) have to be measured in comparison to the three points given above. This is done in the following.

Let $\tilde{\delta}^*$ be the unique solution of (2.22) and $\tilde{r}^* = A\tilde{\delta}^* - b$. With

$$f(\delta) = \|\tilde{A}\delta - \tilde{b}\|_2^2 = \|A\delta - b\|_2^2 + s^2 \sum_{j \in I_B} \delta_j^2 = \|A\delta - b\|_2^2 + s^2 \|\delta_{I_B}\|_2^2 \tag{2.24}$$

we have

$$f\left(\tilde{\delta}^*\right) \leq f\left(\delta^*\right). \tag{2.25}$$

Inequality (2.25) yields for the corresponding residuals

$$\left\|\tilde{r}^*\right\|_2^2 \leq \left\|r^*\right\|_2^2 + s^2\left(\left\|\delta_{I_B}^*\right\|_2^2 - \left\|\tilde{\delta}_{I_B}^*\right\|_2^2\right) \leq \left\|r^*\right\|_2^2 + s^2\left\|\delta_{I_B}^*\right\|_2^2. \tag{2.26}$$

Thus, the maximum deviation of the Euclidean norm of the residual term $\tilde{r}^*$ from the possible minimum $\|r^*\|_2$ is limited and can be reduced by reducing s. In particular, if A is regular (or at least b is in the range of A), we have $\|r^*\|_2 = 0$ and

$$\left\|\tilde{r}^*\right\|_2 \leq s\left\|\delta_{I_B}^*\right\|_2. \tag{2.27}$$

Moreover, we get in the regular case:

$$A\left(\tilde{\delta}^* - \delta^*\right) = \tilde{r}^* \quad \Leftrightarrow \quad \tilde{\delta}^* - \delta^* = A^{-1}\tilde{r}^*. \tag{2.28}$$

Taking the Euclidean norm on both sides yields

$$\left\|\tilde{\delta}^* - \delta^*\right\|_2 \leq \left\|A^{-1}\right\|_2 \left\|\tilde{r}^*\right\|_2 \overset{(2.27)}{\leq} \left\|A^{-1}\right\|_2 s \left\|\delta^*_{I_B}\right\|_2 \leq \left\|A^{-1}\right\|_2 s \left\|\delta^*\right\|_2 \tag{2.29}$$

and finally

$$\frac{\|\tilde{\delta}^* - \delta^*\|_2}{\|\delta^*\|_2} \leq s\left\|A^{-1}\right\|_2 \tag{2.30}$$

for the relative error of $\tilde{\delta}^*$ compared to $\delta^* = A^{-1}b$. Note that $\tilde{\delta}^* = \delta^*$ if $\|\delta^*\|_2 = 0$.

Since $\|\tilde{\delta}^*_{I_B} - \delta^*_{I_B}\|_2 \leq \|\tilde{\delta}^* - \delta^*\|_2$, we also get from (2.29):

$$\frac{\|\tilde{\delta}^*_{I_B} - \delta^*_{I_B}\|_2}{\|\delta^*_{I_B}\|_2} \leq s\left\|A^{-1}\right\|_2. \tag{2.31}$$

Similar to above, note that $\tilde{\delta}^*_{I_B} = \delta^*_{I_B}$ if $\|\delta^*_{I_B}\|_2 = 0$.

In addition to the given results, (2.25) also yields

$$\left\|\tilde{\delta}^*_{I_B}\right\|_2^2 \leq \left\|\delta^*_{I_B}\right\|_2^2 - \frac{1}{s^2}\underbrace{\left(\left\|\tilde{r}^*\right\|_2^2 - \left\|r^*\right\|_2^2\right)}_{\geq 0} \leq \left\|\delta^*_{I_B}\right\|_2^2. \tag{2.32}$$

This means that regarding the indices I_B of the "correction terms" $\tilde{\delta}^*$ and δ^*, the correction induced by the QR decomposition of the modified matrix $\tilde{A}$ is not greater than the one induced by the SVD of the original matrix A. This is an important property since the set of indices I_B typically refers to "critical" variables of the problem, while the rest of the variables is supposed to be determined anyway.

So far, we have given quantitative results for our QR decomposition approach related to the first two advantages of using singular value decomposition. To give a quantitative result related to the third point, we consider the case $I_B = \{1, \ldots, n\}$ with $B_s = sI_n$, where I_n is the n-dimensional identity matrix.

Let $A = U\Sigma V^T$ be a singular value decomposition of A with

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}. \tag{2.33}$$

For the modified matrix $\tilde{A}$ we get

$$\tilde{A}^T\tilde{A} = \begin{pmatrix} A^T & B_s^T \end{pmatrix}\begin{pmatrix} A \\ B_s \end{pmatrix} = A^TA + B_s^2 = V\left(\Sigma^2 + B_s^2\right)V^T. \tag{2.34}$$

Hence, the singular values $\tilde{\sigma}_j$ $(j = 1, \ldots, n)$ of $\tilde{A}$ are given by

$$\tilde{\sigma}_j^2 = \sigma_j^2 + s^2. \tag{2.35}$$

In particular, we have $\tilde{\sigma}_j > \sigma_j$ and $\tilde{\sigma}_j \geq s > 0$.

While in the results above a smaller s is always preferred, here, the opposite is the case since an increase of the singular values leads to more stability. Thus, in practice, a trade-off has to be made.

2.3.3 Theoretical Analysis—Backward Direction

Let $f(y, u)$ be a (scalar) quantity of interest. As described in Sect. 2.2, we can efficiently compute sensitivity information with respect to the control u by solving adjoint equations. Due to the special structure of E, this can also be done time step wise, but backwards in time, and we finally have to solve linear systems of equations with the same matrices as in the forward direction, but transposed.

Thus, we have to solve systems of the form

$$A^T \xi = c. \tag{2.36}$$

In the whole section, we postulate that c is in the range of A^T. This has the following reason: The solution of the simulation process has degrees of freedom in the kernel $\ker(A)$ of A. Thus, regarding the quantity of interest f, it is reasonable to claim that the partial derivatives of f with respect to the state variables (in each time step) are perpendicular to $\ker(A)$, which is equivalent to being in the range of A^T. Additionally to the partial derivatives of f, c also may contain components from the preceding time step. This can only occur in parts of the network where the model contains temporal derivatives, but those parts do not suffer from the described problem of non-uniqueness since the state variables of consecutive time steps are linked here.

Let ξ^* be the solution of (2.36) of minimal Euclidean norm. Similar to Sect. 2.3.2, this can be computed by a singular value decomposition of A respectively A^T. It is natural to apply the QR decomposition of $\tilde{A}$ to solve the modified problem

$$\tilde{A}^T \begin{pmatrix} \xi \\ \mu \end{pmatrix} = c. \tag{2.37}$$

With the QR decomposition

$$\tilde{A} = \begin{pmatrix} \tilde{Q}_{11} & \tilde{Q}_{12} \\ \tilde{Q}_{21} & \tilde{Q}_{22} \end{pmatrix} \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix}, \tag{2.38}$$

where $\tilde{Q}_{11}$ and $\tilde{R}$ are $n \times n$ matrices, the solution of (2.37) of minimal norm can be written as

$$\begin{pmatrix} \tilde{\xi}^* \\ \tilde{\mu}^* \end{pmatrix} = \begin{pmatrix} \tilde{Q}_{11} \\ \tilde{Q}_{21} \end{pmatrix} \tilde{R}^{-T} c. \tag{2.39}$$

This results from the well-known fact that the columns of $\begin{pmatrix} \tilde{Q}_{12} \\ \tilde{Q}_{22} \end{pmatrix}$ form a basis of the kernel of $\tilde{A}^T$. With $\tilde{q}^* = A^T \tilde{\xi}^* - c$ we have

$$\|\tilde{\xi}^*\|_2^2 + \tfrac{1}{s^2}\|\tilde{q}^*\|_2^2 = \|\tilde{\xi}^*\|_2^2 + \|\tilde{\mu}^*\|_2^2 = \left\| \begin{pmatrix} \tilde{\xi}^* \\ \tilde{\mu}^* \end{pmatrix} \right\|_2^2 \leq \left\| \begin{pmatrix} \xi^* \\ 0 \end{pmatrix} \right\|_2^2 = \|\xi^*\|_2^2. \tag{2.40}$$

This yields

$$\|\tilde{q}^*\|_2^2 \leq s^2 \left(\|\xi^*\|_2^2 - \|\tilde{\xi}^*\|_2^2 \right) \leq s^2 \|\xi^*\|_2^2 \tag{2.41}$$

as upper bound for the residual with respect to the original equation (2.36). Analogously to Sect. 2.3.2, we get in the regular case

$$\frac{\|\tilde{\xi}^* - \xi^*\|_2}{\|\xi^*\|_2} \leq s \|A^{-T}\|_2 \tag{2.42}$$

for the relative error of $\tilde{\xi}^*$ compared to $\xi^* = A^{-T} c$.

References

1. KANET. http://kanet.iwg.uni-karlsruhe.de
2. STANET. http://www.stafu.de
3. MATLAB 7.5. The MathWorks Inc. (2007)
4. R. Álvarez, N.B. Gorev, I.F. Kodzhespirova, Y. Kovalenko, S. Negrete, A. Ramos, J.J. Rivera, Pseudotransient continuation method in extended period simulation of water distribution systems. J. Hydraul. Eng. **134**(10), 1473–1479 (2008)
5. R.H. Byrd, J. Nocedal, R.A. Waltz, Knitro: An integrated package for nonlinear optimization, in *Large-Scale Nonlinear Optimization* (2006), pp. 35–59
6. H. Cross, Analysis of flow in networks of conduits or conductors. University of Illinois Bulletin No. 286, 1936
7. T.A. Davis, *Direct Methods for Sparse Linear Systems*, Fundamentals of Algorithms, vol. 2 (Society for Industrial and Applied Mathematics, Philadelphia, 2006)
8. J. Deuerlein, Zur hydraulischen Systemanalyse von Wasserversorgungsnetzen, PhD thesis, U Karlsruhe, 2002
9. J. Deuerlein, A.R. Simpson, E. Gross, The never ending story of modeling control-devices in hydraulic systems analysis, in *Proceedings of Water Distribution Systems Analysis*, ASCE, 2008, p. 72
10. C. Hähnlein, Numerische Modellierung zur Betriebsoptimierung von Wasserverteilnetzen, PhD thesis, TU Darmstadt, 2008
11. O. Kolb, P. Domschke, J. Lang, Modified QR decomposition to avoid non-uniqueness in water supply networks with extension to adjoint calculus. Proc. Comput. Sci. **1**(1), 1421–1428 (2010)

12. O. Kolb, J. Lang, P. Bales, An implicit box scheme for subsonic compressible flow with dissipative source term. Numer. Algorithms **53**(2), 293–307 (2010)
13. S.N. Kružkov, First order quasilinear equations in several independent variables. Math. USSR Sb. **10**(2), 217–243 (1970)
14. R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems* (Cambridge University Press, Cambridge, 2002)
15. L.A. Rossman, *EPANET 2 Users Manual* (U.S. Environmental Protection Agency, Cincinnati, 2000)
16. A.R. Simpson, Modeling of pressure regulating devices: The last major problem to be solved in hydraulic simulation, in *Proceedings of 29th Annual Water Resources Planning and Management Conference*, ASCE, 1999
17. P. Spellucci, A new technique for inconsistent QP problems in the SQP method. Math. Methods Oper. Res. **47**(3), 355–400 (1998)
18. P. Spellucci, An SQP method for general nonlinear programs using only equality constrained subproblems. Math. Program. **82**(3), 413–448 (1998)
19. A. Wächter, L.T. Biegler, On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. Math. Program. **106**(1), 25–57 (2006)

O. Kolb · J. Lang
Numerical Analysis and Scientific Computing, Technische Universität Darmstadt, Dolivostr. 15, 64293 Darmstadt, Germany

O. Kolb
e-mail: kolb@mathematik.tu-darmstadt.de

J. Lang (✉)
e-mail: lang@mathematik.tu-darmstadt.de

http://www.springer.com/978-3-0348-0435-6

Mathematical Optimization of Water Networks

Martin, A.; Klamroth, K.; Lang, J.; Leugering, G.; Morsi, A.;
Oberlack, M.; Ostrowski, M.; Rosen, R. (Eds.)

2012, XIV, 196 p., Hardcover
ISBN: 978-3-0348-0435-6
A product of Birkhäuser Basel