Shallow Water Equations:
Flux-Based Wave Decomposition Solver

Alexander K. Shukaev

June 8, 2013

Abstract

This paper examines numerical solution of the one-dimensional shallow water equations using the *flux-based wave decomposition* approach. Furthermore, the *bathymetry* treatment is also covered in details as it is shown how to seamlessly incorporate the *source term* using this approach. The efficiency of this approach is verified with an open-source implementation which supplements this paper. The theory leading to the development of this approach is addressed briefly as well, but there is extensive referencing to the relevant sources with more information. Finally, the comparison with classical methods such as *Roe linearization* is included.

Keywords

shallow water equations, bathymetry, topography, dam break, hyperbolic partial differential equation, conservation law, balance law, finite volume method, well-balanced method, flux, source term, wave, Riemann problem, Godunov's method

1 Introduction

The bulk of theory comes from [1], a great source on finite volume methods for hyperbolic partial differential equations.

Consider the simple problem of fluid dynamics, in which gas or liquid is flowing through one-dimensional pipe/channel with some known velocity $u(x, t)$, which is assumed to vary only with the point x in the pipe/channel, and time t. Let $\vec{q}(x, t)$ be the *vector of quantity densities*, i.e.

$$\vec{q}(x,t) = \begin{pmatrix} q^{(1)}(x,t) \\ \vdots \\ q^{(p)}(x,t) \\ \vdots \\ q^{(m)}(x,t) \end{pmatrix}, \tag{1.1}$$

where $q^{(p)}(x, t)$ for $p = \{1, ..., m\}$ is the density of p-th quantity at the point x in the pipe/channel, and time t. Since this is the one-dimensional case, it is natural to assume that $q^{(p)}$ is measured in *units of quantity per length*. For instance, taking fluid density ρ as a quantity density of interest means that it could be measured in *grams per meter*. As a result,

$$\int\limits_{x_1}^{x_2} \vec{q}(x,t)\,dx \tag{1.2}$$

is the vector each p-th component of which represents the *total amount* of the corresponding quantity between x_1 and x_2 at the particular moment in time t.

If we are concerned with a substance whose quantities can neither be created nor destroyed within the section (x_1, x_2), then the total amount of quantities within this section can only change due to the *flux* of these quantities through the endpoints of the section at x_1 and x_2. For a general autonomous flux $\vec{f}(\vec{q})$, that depends only on the vector $\vec{q}$, the *conservation law* can be written as

$$\frac{d}{dt}\int\limits_{x_1}^{x_2} \vec{q}(x,t)\,dx = -\vec{f}(\vec{q}(x,t))\Big|_{x_1}^{x_2}. \tag{1.3}$$

If we assume that both $\vec{q}$ and $\vec{f}$ have smooth components, then the equation (1.3) can be rewritten as

$$\frac{d}{dt}\int\limits_{x_1}^{x_2} \vec{q}(x,t)\,dx = -\int\limits_{x_1}^{x_2} \frac{\partial}{\partial x}\vec{f}(\vec{q}(x,t))\,dx,$$

which could be rearranged to

$$\int\limits_{x_1}^{x_2} \left(\frac{\partial}{\partial t}\vec{q}(x,t) + \frac{\partial}{\partial x}\vec{f}(\vec{q}(x,t))\right)\,dx = \vec{0}.$$

We can now recall that the interval $[x_1, x_2]$ was chosen arbitrarily, and this implies that the integrand must be identically zero everywhere within $[x_1, x_2]$. If this were not so (e.g. the integrand is zero because there are positive and negative contributions that cancel out), then we could subdivide $[x_1, x_2]$ into smaller subintervals over each of which the integral would be nonzero, and hence violating the fact that it should actually be zero. Thus, we can conclude that

$$\frac{\partial}{\partial t}\vec{q}(x,t) + \frac{\partial}{\partial x}\vec{f}(\vec{q}(x,t)) \equiv \vec{0}. \tag{1.4}$$

This is known as the *differential form* of the conservation law. Again, it should be stressed that this differential form of the conservation law is derived from the more fundamental *integral form* (1.3) under the assumption that both $\vec{q}$ and $\vec{f}$ are smooth enough, which might be not true in certain cases. Furthermore, when $\vec{q}$ is smooth, it is possible to rewrite (1.4) as

$$\frac{\partial \vec{q}}{\partial t} + \vec{f}'(\vec{q}) \cdot \frac{\partial \vec{q}}{\partial x} = \vec{0}, \tag{1.5}$$

where $\vec{f}'(\vec{q})$ is the derivative of a vector with respect to a vector, which is clearly the $m \times m$ Jacobian matrix. Equation (1.5) is called the *quasilinear form* as it resembles the linear system

$$\frac{\partial \vec{q}}{\partial t} + A \cdot \frac{\partial \vec{q}}{\partial x} = \vec{0}, \tag{1.6}$$

where A is $m \times m$ matrix independent of $\vec{q}$.

Definition 1.1 *The $m \times m$ matrix A is* diagonalizable *if there is a* complete *set of linearly independent eigenvectors $\vec{r}^{(p)} \neq \vec{0}$ such that*

$$A \cdot \vec{r}^{(p)} = \lambda^{(p)} \cdot \vec{r}^{(p)}$$

for $p = \{1, ..., m\}$.

Since eigenvectors $\vec{r}^{(p)}$ are linearly independent, it is possible to form a matrix

$$R = \begin{pmatrix} \vec{r}^{(1)} & \cdots & \vec{r}^{(p)} & \cdots & \vec{r}^{(m)} \end{pmatrix},$$

which is nonsingular, and therefore has an inverse R^{-1}. Thus, matrix A can be factorized in a form of

$$A = R \cdot \Lambda \cdot R^{-1}, \tag{1.7}$$

where

$$\Lambda = \begin{pmatrix} \lambda^{(1)} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda^{(p)} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \lambda^{(m)} \end{pmatrix}.$$

Definition 1.2 *Linear system of the form (1.6) is called* hyperbolic *if the $m \times m$ matrix A is diagonalizable with real eigenvalues.*

As a result, we can use (1.7) to rewrite (1.6) as

$$\frac{\partial \overrightarrow{w}}{\partial t} + \Lambda \cdot \frac{\partial \overrightarrow{w}}{\partial x} = \vec{0}, \tag{1.8}$$

where $\overrightarrow{w}(x,t) = R^{-1} \cdot \vec{q}(x,t)$.

Since Λ is diagonal matrix, equation (1.8) represents the system of m *decoupled advection equations* for the respective components $w^{(p)}$ of $\overrightarrow{w}$:

$$\frac{\partial w^{(p)}}{\partial t} + \lambda^{(p)} \cdot \frac{\partial w^{(p)}}{\partial x} = 0 \tag{1.9}$$

for $p = \{1, ..., m\}$.

2 The Cauchy Problem

Consider the *Cauchy problem* (see [1] for more details) for the constant-coefficient system (1.6), for which we are given the initial distribution vector of quantity densities $\vec{q}(x, t = 0)$ for $x \in \mathbb{R}$. From this data we can compute

$$\overrightarrow{w}(x,0) = R^{-1} \cdot \vec{q}(x,0)$$

for the system (1.8). Accordingly, p-th advection equation (1.9) has a solution

$$w^{(p)}(x,t) = w^{(p)}(x - \lambda^{(p)} \cdot t, 0). \tag{2.1}$$

The solution to the original problem (1.6) is obtained by combining all the components $w^{(p)}(x,t)$ into the vector $\overrightarrow{w}(x,t)$ and performing the inverse transformation

$$\vec{q}(x,t) = R \cdot \overrightarrow{w}(x,t) = \sum_{p=1}^{m} w^{(p)}(x,t) \cdot \vec{r}^{(p)}. \tag{2.2}$$

The emphasis is put on the fact that the solution $\vec{q}(x,t)$ to the original problem (1.6) is being a linear combination of the right eigenvectors $\vec{r}^{(1)}, ..., \vec{r}^{(m)}$ at each point in space and time. Hence, it is essentially a superposition of m independent waves propagating at the corresponding velocities $\lambda^{(p)}$. Furthermore, the structure of p-th advection equation (1.9) suggests that each initial profile $w^{(p)}(x,0)$ is simply *advected* at a constant *characteristic speed* $\lambda^{(p)}$ as time evolves (**Figure 2.1**), i.e. $w^{(p)}(x,t) \equiv w^{(p)}(x_0, 0)$ all along the ray $x(t) = x_0 + \lambda^{(p)} \cdot t$. These rays are called *p-characteristics*.

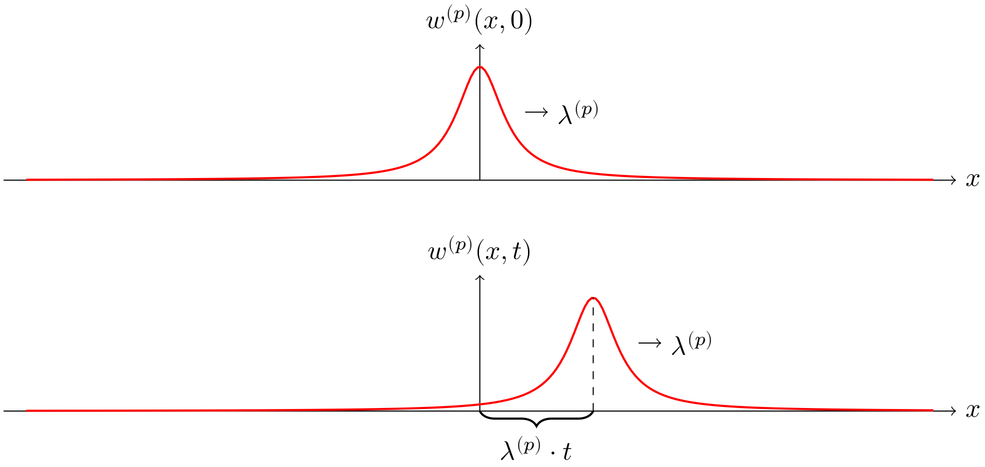

Figure 2.1 Advection of initial profile $w^{(p)}(x,0)$ at constant speed $\lambda^{(p)}$

3 The Linear Riemann Problem

While classical solutions of differential equations must be smooth (sufficiently differentiable) functions, the formula **(2.2)** can be used even if the initial profile of quantity densities $\vec{q}(x,0)$ is not smooth, or is even discontinuous, at some points (refer to [1] for more information on that).

If the initial profile has singularity (a discontinuity in some derivative) at some point x_0, then one or more of the *characteristic variables* $w^{(p)}(x,0)$ will also have a singularity at this point. According to [1], such singularities can then propagate along the corresponding p-characteristics and lead to singularities in the solution $\vec{q}(x,t)$ at some or all of the points $x(t) = x_0 + \lambda^{(p)} \cdot t$.

The linear *Riemann problem* consists of the linear hyperbolic equation **(1.6)** together with special initial profile of quantity densities $\vec{q}(x,0)$ that is piecewise constant with a single jump discontinuity

$$\vec{q}(x,0) = \begin{cases} \vec{q}_l, \text{ if } x < 0 \\ \vec{q}_r, \text{ if } x > 0 \end{cases}.$$

It is natural to decompose both $\vec{q}_l$ and $\vec{q}_r$ as

$$\vec{q}_l = \sum_{p=1}^{m} w_l^{(p)} \cdot \vec{r}^{(p)} \tag{3.1}$$

and

$$\vec{q}_r = \sum_{p=1}^{m} w_r^{(p)} \cdot \vec{r}^{(p)}. \tag{3.2}$$

Then the initial profile of the characteristic variable in p-th advection equation **(1.9)** is

$$w^{(p)}(x,0) = \begin{cases} w_l^{(p)}, \text{ if } x < 0 \\ w_r^{(p)}, \text{ if } x > 0 \end{cases}, \tag{3.3}$$

and according to **(2.1)** this discontinuity propagates with the speed $\lambda^{(p)}$:

$$w^{(p)}(x,t) = \begin{cases} w_l^{(p)}, \text{ if } x - \lambda^{(p)} \cdot t < 0 \\ w_r^{(p)}, \text{ if } x - \lambda^{(p)} \cdot t > 0 \end{cases}. \tag{3.4}$$

As a result, by taking into consideration **(2.2)**, it makes sense to write

$$\vec{q}(x,t) = \sum_{p:\ x > \lambda^{(p)} \cdot t} w_r^{(p)} \cdot \vec{r}^{(p)} + \sum_{p:\ x < \lambda^{(p)} \cdot t} w_l^{(p)} \cdot \vec{r}^{(p)}. \tag{3.5}$$

The solution $\vec{q}(x,t)$ is constant in each of the wedges as shown in **Figure 3.1**. Across the p-th characteristic the solution jump is given by

$$(w_r^{(p)} - w_l^{(p)}) \cdot \vec{r}^{(p)} = \alpha^{(p)} \cdot \vec{r}^{(p)} = \overrightarrow{\mathcal{W}}^{(p)}, \tag{3.6}$$

where $\overrightarrow{\mathcal{W}}^{(p)}$ will be further called *wave*. The interesting observation is that wave $\overrightarrow{\mathcal{W}}^{(p)}$ is an eigenvector of the matrix A. This is known as the *Rankine-Hugoniot condition* (see [1]), and is extremely important as it allows to solve the Riemann problem for nonlinear problems properly (discussed in **Section 7**).

For instance, if there are two equations in the system **(1.6)**, then the solution to the linear Riemann problem has three states (**Figure 3.1**):

- the original state to the left $\vec{q}_l$;
- the original state to the right $\vec{q}_r$;
- the middle state $\vec{q}_*$ between the two discontinuities (waves) $\overrightarrow{\mathcal{W}}^{(1)}$ and $\overrightarrow{\mathcal{W}}^{(2)}$.

For the case of a linear system **(1.6)**, solving the Riemann problem consists of decomposing the initial jump $\vec{q}_r - \vec{q}_l$ into eigenvectors of the matrix A:

$$\vec{q}_r - \vec{q}_l = \sum_{p=1}^{m} \alpha^{(p)} \cdot \vec{r}^{(p)}. \tag{3.7}$$

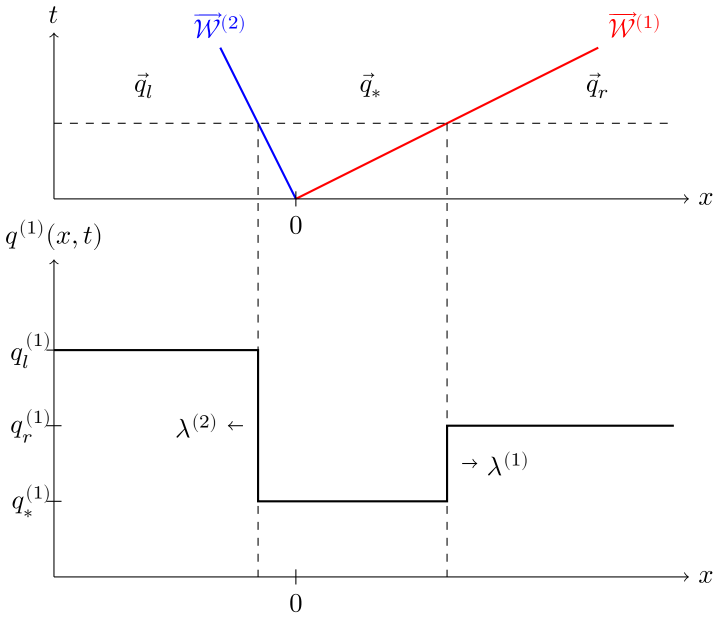

Figure 3.1 An example of the solution to the linear Riemann problem for $m = 2$

Essentially, this requires solving the linear system of equations

$$R \cdot \vec{\alpha} = \vec{q}_r - \vec{q}_l \tag{3.8}$$

for the vector

$$\vec{\alpha} = \begin{pmatrix} \alpha^{(1)} \\ \vdots \\ \alpha^{(p)} \\ \vdots \\ \alpha^{(m)} \end{pmatrix}.$$

Finally, by combining (3.1) and (3.2) with (3.5), the solution $\vec{q}(x,t)$ can be rewritten in terms of the waves in two equivalent forms:

$$\vec{q}(x,t) = \vec{q}_l + \sum_{p:\ x>\lambda^{(p)}\cdot t} \vec{W}^{(p)} \tag{3.9}$$

and

$$\vec{q}(x,t) = \vec{q}_r - \sum_{p:\ x<\lambda^{(p)}\cdot t} \vec{W}^{(p)}. \tag{3.10}$$

More details on the Riemann problem can be found in [1].

4 Finite Volume Methods

Finite volume methods are derived on the basis of the integral form of the conservation law **(1.3)**, and this is the main point where they gain their advantages over naive numerical approaches to solve hyperbolic problems (such as *finite difference methods*).

From the numerical standpoint, *shock waves* introduce major difficulties and prevent one from utilizing quite some methods (see [1]). For instance, application of naive methods could easily lead to the following consequences:

- drastic oscillations around shock waves;
- propagation of shock waves at wrong speeds;
- computation of shock waves with wrong strengths.

According to [1], finite volume methods are well-suited for hyperbolic problems which allow discontinuities (shock waves) in the *weak solution*. At the same time finite volume methods can approximate smooth parts (such as *rarefaction wave*) of the solutions very well.

In one space dimension, a finite volume method is based on subdividing the spatial domain into intervals (the *finite volumes*, also called *grid cells*) and keeping track of an approximation to the integral of $\vec{q}$ over each of these finite volumes. Denote the i-th grid cell by

$$\mathcal{C}_i = (x_{i-1/2}, x_{i+1/2}).$$

As described in [1], picking the spatial integral in **(1.3)** over the grid cell $\mathcal{C}_i$, integrating over a time interval $[t_n, t_{n+1}]$, and dividing by Δx in order to obtain grid cell averages of $\vec{q}$ gives

$$\begin{aligned}\frac{1}{\Delta x}\int\limits_{\mathcal{C}_i} \vec{q}(x, t_{n+1})\,dx &= \frac{1}{\Delta x}\int\limits_{\mathcal{C}_i} \vec{q}(x, t_n)\,dx - \\ &\quad \frac{1}{\Delta x}\cdot\left(\int\limits_{t_n}^{t_{n+1}} \vec{f}(\vec{q}(x_{i+1/2}, t))\,dt - \int\limits_{t_n}^{t_{n+1}} \vec{f}(\vec{q}(x_{i-1/2}, t))\,dt\right).\end{aligned}$$

This can be compactly rewritten as

$$\vec{Q}_i^{n+1} = \vec{Q}_i^n - \frac{\Delta t}{\Delta x}\cdot\left(\vec{F}_{i+1/2}^n - \vec{F}_{i-1/2}^n\right), \tag{4.1}$$

where

$$\vec{Q}_i^n \approx \frac{1}{\Delta x} \int\limits_{\mathcal{C}_i} \vec{q}(x, t_n)\, dx$$

and

$$\vec{Q}_i^{n+1} \approx \frac{1}{\Delta x} \int\limits_{\mathcal{C}_i} \vec{q}(x, t_{n+1})\, dx$$

are the spatial averages of $\vec{q}$ over the grid cell $\mathcal{C}_i$ at times t_n and t_{n+1} respectively, and

$$\vec{F}_{i\pm1/2}^n \approx \frac{1}{\Delta t} \int\limits_{t_n}^{t_{n+1}} \vec{f}(\vec{q}(x_{i\pm1/2}, t))\, dt$$

are the temporary averages of fluxes through boundaries of the grid cell $\mathcal{C}_i$ over a time interval $[t_n, t_{n+1}]$.

Of course the variation of $\vec{q}(x_{i\pm1/2}, t)$ in time t is unknown. Therefore $\vec{F}_{i\pm1/2}^n$ cannot be evaluated directly, but should rather be reasonably approximated. For hyperbolic problems information propagates with finite speed, so it makes sense to suppose that it could be possible to obtain $\vec{F}_{i-1/2}^n$, for example, based only on $\vec{Q}_{i-1}^n$ and $\vec{Q}_i^n$ (which are the grid cell averages on either side of this interface). This idea can be expressed in the following form:

$$\vec{F}_{i-1/2}^n = \vec{\mathcal{F}}(\vec{Q}_{i-1}^n, \vec{Q}_i^n) \tag{4.2}$$

and

$$\vec{F}_{i+1/2}^n = \vec{\mathcal{F}}(\vec{Q}_i^n, \vec{Q}_{i+1}^n), \tag{4.3}$$

where $\vec{\mathcal{F}}$ is called *numerical flux*.

Melting (4.1) and (4.2)/(4.3) together essentially defines the class of finite volume methods. The specific method obtained depends exclusively on the choice of formula for $\vec{\mathcal{F}}$.

In any case, the most important property of all finite volume methods (4.1) is that they are *conservative* as they conserve the numerical variable $\vec{Q}$ on the whole computational domain. In other words, they behave in the same way as the integral form of the conservation law (1.3) does, which could be proved through summation of $\Delta x \cdot \vec{Q}_i^{n+1}$ over an arbitrary sequence of adjacent grid cells:

$$\Delta x \cdot \sum_{i=i_1}^{i_2} \vec{Q}_i^{n+1} = \Delta x \cdot \sum_{i=i_1}^{i_2} \vec{Q}_i^{n} - \frac{\Delta t}{\Delta x} \cdot \left(\vec{F}_{i_2+1/2}^{n} - \vec{F}_{i_1-1/2}^{n} \right) . \tag{4.4}$$

The sum of the numerical flux differences cancels out, except for the fluxes at the very edges of the target domain, in the exact same way as it would for (1.3).

Furthermore, when numerically approximating integral conservation law (1.3) with discontinuous weak solutions, conservative methods based on the integral conservation law (1.3) (such as finite volume methods) converge properly to a discontinuous weak solution, whereas methods based on the differential conservation law (1.4) (such as finite difference methods) alone often fail to converge to a correct weak solution. Actually, [5] provides a proof that nonconservative methods converge to wrong solutions if the real solution is supposed to contain shock waves. Refer to [1] for more information on this topic.

5 Godunov's Method

As discussed earlier, development of a conservative flux-differencing method based on (4.1) merely boils down to establishing a proper estimate for the temporary averages of fluxes (4.2) and (4.3). *Godunov's method* has proved to be successful at dealing with many of the difficulties of hyperbolic conservation laws, including shock waves. Godunov's method is *upwind* method where the Riemann problem is solved for left and right interfaces of each grid cell $\mathcal{C}_i$ on every time step (see [1]).

The reason for this is that from the numerical point of view, the solution is basically a piecewise constant function with the grid cell averages $\vec{Q}_i$ at each grid cell $\mathcal{C}_i$. Thus, there are discontinuities between each pair of neighboring states $\vec{Q}_{i-1}$ and $\vec{Q}_i$ on the left, and $\vec{Q}_i$ and $\vec{Q}_{i+1}$ on the right at the grid cell interfaces $x_{i-1/2}$ and $x_{i+1/2}$ respectively. In Godunov's method the numerical fluxes (4.2) and (4.3) are determined by evaluation of the true flux function $\vec{f}(\vec{q})$ with the solution to the Riemann problem at each grid cell interface:

$$\vec{Q}_i^{n+1} = \vec{Q}_i^{n} - \frac{\Delta t}{\Delta x} \cdot \left(\vec{f}(\vec{Q}_{i+1/2}^{n}) - \vec{f}(\vec{Q}_{i-1/2}^{n}) \right) , \tag{5.1}$$

where $\vec{Q}_{i\pm1/2}^{n}$ are the solutions to the left and right Riemann problems (respective to the i-th grid cell) at the corresponding grid cell interfaces $x_{i\pm1/2}$.

Notice that in the case of the linear Riemann problem (see **Section 3**) $\vec{Q}^n_{i\pm1/2}$ are merely middle states $\vec{q}_*$ (see **Figure 3.1**) of the solutions to the left and right linear Riemann problems respectively. This fact is very important as it greatly simplifies the numerical solution of nonlinear hyperbolic problems by local linearization of the problem utilizing the quasilinear formulation **(1.5)** and, as a result, solving the linear Riemann problem instead the nonlinear one (see **Section 7**).

According to [1], for a general conservation law **(1.4)**, utilizing **(3.9)** or **(3.10)**, left numerical flux **(4.2)** can be defined as

$$\vec{F}^n_{i-1/2} = \vec{f}(\vec{Q}_{i-1}) + \sum_{p=1}^{m} (\lambda^{(p)}_{i-1/2})^- \cdot \vec{W}^{(p)}_{i-1/2} \tag{5.2}$$

or

$$\vec{F}^n_{i-1/2} = \vec{f}(\vec{Q}_{i}) - \sum_{p=1}^{m} (\lambda^{(p)}_{i-1/2})^+ \cdot \vec{W}^{(p)}_{i-1/2}, \tag{5.3}$$

and right numerical flux **(4.3)** can be defined as

$$\vec{F}^n_{i+1/2} = \vec{f}(\vec{Q}_{i}) + \sum_{p=1}^{m} (\lambda^{(p)}_{i+1/2})^- \cdot \vec{W}^{(p)}_{i+1/2} \tag{5.4}$$

or

$$\vec{F}^n_{i+1/2} = \vec{f}(\vec{Q}_{i+1}) - \sum_{p=1}^{m} (\lambda^{(p)}_{i+1/2})^+ \cdot \vec{W}^{(p)}_{i+1/2}, \tag{5.5}$$

where

$$(\lambda)^+ = \max(\lambda, 0)$$

and

$$(\lambda)^- = \min(\lambda, 0).$$

Substituting **(5.3)** and **(5.4)** into **(4.1)**, we arrive to the most common implementation of Godunov's method:

$$\vec{Q}^{n+1}_i = \vec{Q}^n_i - \frac{\Delta t}{\Delta x} \cdot \left(\sum_{p=1}^{m} (\lambda^{(p)}_{i+1/2})^- \cdot \vec{W}^{(p)}_{i+1/2} + \sum_{p=1}^{m} (\lambda^{(p)}_{i-1/2})^+ \cdot \vec{W}^{(p)}_{i-1/2} \right). \tag{5.6}$$

For a more thorough discussion of this topic refer to [1].

6 Shallow Water Equations

To derive the one-dimensional *shallow water equations*, we consider fluid in a channel of unit width and assume that the vertical velocity of the fluid is negligible, while the horizontal velocity $u(x,t)$ is roughly constant throughout any cross section of the channel. This is true if we consider small-amplitude waves in the fluid that is shallow relative to the wavelength.

Let the fluid be incompressible, so the density ρ is constant. The height $h(x,t)$ of the fluid surface can vary. Thus, according to **(1.2)** the total mass in $[x_1, x_2]$ at time t is

$$\int\limits_{x_1}^{x_2} \rho \cdot h(x,t)\, dx.$$

The mass flux is $\rho \cdot h(x,t) \cdot u(x,t)$, so considering **(1.4)** and dropping ρ we can express the *conservation of mass* as

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(h \cdot u) = 0. \tag{6.1}$$

In addition to this equation we now need a second equation for the velocity. The velocity itself is not a conserved quantity, but the momentum is. The product $\rho \cdot u(x,t)$ gives the density of momentum, in the sense that

$$\int\limits_{x_1}^{x_2} \rho \cdot u(x,t)\, dx$$

yields the total momentum in the interval $[x_1, x_2]$, and this can change only due to the flux of momentum through the endpoints x_1 and x_2 of the interval. The momentum flux past any point x consists of two terms:

- *convection* – advection of the density of momentum: $\rho \cdot u^2$;
- microscopic momentum flux due to *pressure* of the fluid: p.

The pressure p can be determined from a well-known *hydrostatic law*

$$p = \frac{1}{2} \cdot \rho \cdot g \cdot h^2, \tag{6.2}$$

where g is gravitational acceleration on the surface of the Earth.

Finally, utilizing **(1.4)** and dropping ρ again, we obtain the equation for the *conservation of momentum* in the form

$$\frac{\partial}{\partial t}(h \cdot u) + \frac{\partial}{\partial x}(h \cdot u^2 + \frac{1}{2} \cdot g \cdot h^2) = 0. \tag{6.3}$$

We can combine equations **(6.1)** and **(6.3)** into the system of one-dimensional shallow water equations

$$\frac{\partial}{\partial t}\begin{pmatrix} h \\ h \cdot u \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} h \cdot u \\ h \cdot u^2 + \frac{1}{2} \cdot g \cdot h^2 \end{pmatrix} = \vec{0}. \tag{6.4}$$

Further discussion of shallow water equations can be found in [1].

7 The Flux-Based Wave Decomposition Approach

As discussed in **Section 5**, the classical implementation of Godunov's method consists of splitting the jump in $\vec{q}$ between two adjacent grid cell averages into waves as in **(3.7)**:

$$\vec{Q}_i - \vec{Q}_{i-1} = \sum_{p=1}^{m} \alpha^{(p)}_{i-1/2} \cdot \hat{\vec{r}}^{(p)}_{i-1/2} = \sum_{p=1}^{m} \vec{W}^{(p)}_{i-1/2} \tag{7.1}$$

and

$$\vec{Q}_{i+1} - \vec{Q}_i = \sum_{p=1}^{m} \alpha^{(p)}_{i+1/2} \cdot \hat{\vec{r}}^{(p)}_{i+1/2} = \sum_{p=1}^{m} \vec{W}^{(p)}_{i+1/2}, \tag{7.2}$$

where $\hat{\vec{r}}^{p}_{i\pm1/2}$ are eigenvectors of averaged Jacobians $\hat{A}_{i\pm1/2}$ respectively. Both waves $\vec{W}^{(p)}_{i\pm1/2}$ are propagating at the corresponding constant speeds $\hat{\lambda}^{(p)}_{i\pm1/2}$, which are eigenvalues of averaged Jacobians $\hat{A}_{i\pm1/2}$ respectively (see **Section 2** and **Section 3**).

For nonlinear problems such as shallow water equations **(6.4)**, where solutions to the nonlinear Riemann problem containing rarefaction wave are possible, in order to obtain conservative finite volume method, the conditions

$$\hat{A}_{i-1/2} \cdot (\vec{Q}_i - \vec{Q}_{i-1}) = \vec{f}(\vec{Q}_i) - \vec{f}(\vec{Q}_{i-1}) \tag{7.3}$$

and

$$\hat{A}_{i+1/2} \cdot (\vec{Q}_{i+1} - \vec{Q}_i) = \vec{f}(\vec{Q}_{i+1}) - \vec{f}(\vec{Q}_i) \tag{7.4}$$

have to be imposed on both Jacobian averages $\hat{A}_{i\pm1/2}$ respectively. Typically, this is achieved with Roe linearization (see [1] for more details). As a result, Godunov's method can be implemented accordingly to **(5.6)**:

$$\overrightarrow{Q}_i^{n+1} = \overrightarrow{Q}_i^n - \frac{\Delta t}{\Delta x} \cdot \left(\sum_{p=1}^{m} (\hat{\lambda}_{i+1/2}^{(p)})^- \cdot \overrightarrow{W}_{i+1/2}^{(p)} + \sum_{p=1}^{m} (\hat{\lambda}_{i-1/2}^{(p)})^+ \cdot \overrightarrow{W}_{i-1/2}^{(p)} \right). \tag{7.5}$$

Otherwise, **(7.5)** cannot be used. Considerable amount of effort has been put into defining properly averaged Jacobians having property **(7.3)** and **(7.4)** for various nonlinear problems such as shallow water equations **(6.4)** or Euler equations for gas dynamics (see [1]).

In contrast, the numerical method presented in this paper is the conservative finite volume method based on *flux-based wave decomposition* approach. It solves the problem by taking the different point of view, what results in several significant advantages over the classical **(7.5)** described above. It was first formulated in [2], and studied in more detail in [3]. General discussions can be found in [1] as well.

Considering the linearized Riemann problem (see **Section 3**) on grid cell interfaces again, the idea is to decompose the flux difference $\vec{f}(\overrightarrow{Q}_i) - \vec{f}(\overrightarrow{Q}_{i-1})$ into a linear combination of the eigenvectors $\hat{\vec{r}}_{i-1/2}^{p}$ of the averaged Jacobian $\hat{A}_{i-1/2}$. As a result, instead of solving the systems **(7.1)** and **(7.2)**, it is possible to solve

$$\vec{f}(\overrightarrow{Q}_i) - \vec{f}(\overrightarrow{Q}_{i-1}) = \sum_{p=1}^{m} \beta_{i-1/2}^{(p)} \cdot \hat{\vec{r}}_{i-1/2}^{(p)} \tag{7.6}$$

and

$$\vec{f}(\overrightarrow{Q}_{i+1}) - \vec{f}(\overrightarrow{Q}_i) = \sum_{p=1}^{m} \beta_{i+1/2}^{(p)} \cdot \hat{\vec{r}}_{i+1/2}^{(p)} \tag{7.7}$$

for the coefficients $\beta_{i\pm1/2}^{(p)}$ respectively, and then define *flux waves* as

$$\overrightarrow{\mathcal{Z}}_{i-1/2}^{(p)} = \beta_{i-1/2}^{(p)} \cdot \hat{\vec{r}}_{i-1/2}^{(p)} \tag{7.8}$$

and

$$\overrightarrow{\mathcal{Z}}_{i+1/2}^{(p)} = \beta_{i+1/2}^{(p)} \cdot \hat{\vec{r}}_{i+1/2}^{(p)}. \tag{7.9}$$

The Godunov's method implementation then looks as follows

$$\vec{Q}_i^{n+1} = \vec{Q}_i^n - \frac{\Delta t}{\Delta x} \cdot \left(\sum_{p=1}^{m} \left|\operatorname{sgn}^-(\hat{\lambda}_{i+1/2}^{(p)})\right| \cdot \vec{\mathcal{Z}}_{i+1/2}^{(p)} + \sum_{p=1}^{m} \left|\operatorname{sgn}^+(\hat{\lambda}_{i-1/2}^{(p)})\right| \cdot \vec{\mathcal{Z}}_{i-1/2}^{(p)} \right), \tag{7.10}$$

where

$$\operatorname{sgn}^+(\lambda) = \max(\operatorname{sgn}(\lambda), 0)$$

and

$$\operatorname{sgn}^-(\lambda) = \min(\operatorname{sgn}(\lambda), 0).$$

According to [1] and [2], the first potential advantage of the decomposition (7.6) over (7.1) is that the resulting method (7.10) is conservative even if (7.3) is not satisfied, whereas, as stated above, (7.5) is not. For instance, even the simple arithmetically averaged Jacobians

$$\hat{A}_{i-1/2} = \vec{f}' \left(\frac{\vec{Q}_i + \vec{Q}_{i-1}}{2} \right) \tag{7.11}$$

and

$$\hat{A}_{i+1/2} = \vec{f}' \left(\frac{\vec{Q}_{i+1} + \vec{Q}_i}{2} \right) \tag{7.12}$$

could be used in case of (7.10), while for (7.5) they couldn't. This yields a more flexible and generic algorithm for hyperbolic problems where Roe averaged Jacobians are either not easily computed or couldn't be defined at all.

However, to emphasize the relation between the novel (7.10) and the classical (7.5), imagine that both Jacobian approximations $\hat{A}_{i\pm1/2}$ do in fact satisfy (7.3) and (7.4) respectively, then multiplying (7.1) by $\hat{A}_{i-1/2}$ yields

$$\vec{f}(\vec{Q}_i) - \vec{f}(\vec{Q}_{i-1}) = \sum_{p=1}^{m} \alpha_{i-1/2}^{(p)} \cdot \hat{\lambda}_{i-1/2}^{(p)} \cdot \hat{\vec{r}}_{i-1/2}^{(p)} \tag{7.13}$$

and multiplying (7.2) by $\hat{A}_{i+1/2}$ yields

$$\vec{f}(\vec{Q}_{i+1}) - \vec{f}(\vec{Q}_i) = \sum_{p=1}^{m} \alpha_{i+1/2}^{(p)} \cdot \hat{\lambda}_{i+1/2}^{(p)} \cdot \hat{\vec{r}}_{i+1/2}^{(p)}. \tag{7.14}$$

That suggests the following relations between quantity waves and flux waves:

$$\vec{\mathcal{Z}}^{(p)}_{i-1/2} = \hat{\lambda}^{(p)}_{i-1/2} \cdot \vec{\mathcal{W}}^{(p)}_{i-1/2} \tag{7.15}$$

and

$$\vec{\mathcal{Z}}^{(p)}_{i+1/2} = \hat{\lambda}^{(p)}_{i+1/2} \cdot \vec{\mathcal{W}}^{(p)}_{i+1/2}. \tag{7.16}$$

8 Source Term Treatment

The *differential balance law*

$$\frac{\partial}{\partial t}\vec{q}(x,t) + \frac{\partial}{\partial x}\vec{f}(\vec{q}(x,t)) = \vec{\psi}(\vec{q}(x,t),x) \tag{8.1}$$

consists of a differential conservation law **(1.4)** with a source term $\vec{\psi}(\vec{q}(x,t),x)$ on the right-hand side.

A method is called *well-balanced* if equilibrium initial data is exactly preserved by the method. Furthermore, the method should also accurately resolve solutions that are small deviations from equilibrium data.

Many numerical approaches have been studied to solve **(8.1)** properly. For example, the so-called *fractional-step method* discussed in [1], which boils down to switching between solving **(1.4)** and

$$\frac{\partial}{\partial t}\vec{q}(x,t) = \vec{\psi}(\vec{q}(x,t),x).$$

However, this approach does not work well if the solution is close to equilibrium state (see [4]), i.e. when

$$\frac{\partial}{\partial x}\vec{f}(\vec{q}(x,t)) \approx \vec{\psi}(\vec{q}(x,t),x), \tag{8.2}$$

while each term separately is large.

A better method was proposed in [2], and studied deeper in [3]. It is again based on the flux-based wave decomposition presented in **Section 7**. In fact, another advantage of the flux-based wave decomposition, which is of particular interest here, is that one can incorporate source terms directly into **(7.6)** and **(7.7)** to get

$$\vec{f}(\vec{Q}_i) - \vec{f}(\vec{Q}_{i-1}) - \Delta x \cdot \vec{\Psi}_{i-1/2} = \sum_{p=1}^{m} \vec{\mathcal{Z}}^{(p)}_{i-1/2} \tag{8.3}$$

and

$$\vec{f}(\vec{Q}_{i+1}) - \vec{f}(\vec{Q}_i) - \Delta x \cdot \vec{\Psi}_{i+1/2} = \sum_{p=1}^{m} \vec{\mathcal{Z}}^{(p)}_{i+1/2}. \tag{8.4}$$

In contrast to the fractional-step method, this method is particularly attractive in cases where the solution is close to the equilibrium state, in which **(8.2)** takes place. To understand why, recall **(8.2)** and consider left and right discretized versions of it:

$$\frac{\vec{f}(\vec{Q}_i) - \vec{f}(\vec{Q}_{i-1})}{\Delta x} = \vec{\Psi}_{i-1/2} \tag{8.5}$$

and

$$\frac{\vec{f}(\vec{Q}_{i+1}) - \vec{f}(\vec{Q}_i)}{\Delta x} = \vec{\Psi}_{i+1/2}, \tag{8.6}$$

then the left-hand sides of both **(8.3)** and **(8.4)** will be zero respectively, and hence all the flux waves will have zero strength, which is the indication of numerical equilibrium state satisfying **(8.5)** and **(8.6)** being maintained exactly. As a result, it turns out that the method can be well-balanced. The only trick is to choose an appropriate averaging scheme for the source term.

Underwater *topography* is generally called *bathymetry* further denoted by $B(x)$ (see **Figure 8.1**). The free surface of the fluid is then given by

$$S(x, t) = h(x, t) + B(x). \tag{8.7}$$

Furthermore, $H(x)$ denotes the distance from bathymetry $B(x)$ to some constant *reference level* L of the fluid surface (see **Figure 8.1**):

$$L = H(x) + B(x). \tag{8.8}$$

The shallow water equations **(6.4)** then take the form

$$\frac{\partial}{\partial t}\begin{pmatrix} h \\ h \cdot u \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} h \cdot u \\ h \cdot u^2 + \dfrac{1}{2} \cdot g \cdot h^2 \end{pmatrix} = \begin{pmatrix} 0 \\ g \cdot h \cdot H'(x) \end{pmatrix}, \tag{8.9}$$

but by applying **(8.8)**, it is convenient to rewrite them as

$$\frac{\partial}{\partial t}\begin{pmatrix} h \\ h \cdot u \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} h \cdot u \\ h \cdot u^2 + \dfrac{1}{2} \cdot g \cdot h^2 \end{pmatrix} = \begin{pmatrix} 0 \\ -g \cdot h \cdot B'(x) \end{pmatrix}, \tag{8.10}$$

where, according to (8.1),

$$\vec{\psi}(\vec{q}, x) = \begin{pmatrix} 0 \\ -g \cdot q^{(1)} \cdot B'(x) \end{pmatrix}. \tag{8.11}$$

As discussed in [3], the so-called *surface-at-rest* is the important equilibrium case arising in many applications, and it can be expressed as

$$\begin{cases} u_e \equiv 0 \\ h_e(x) + B(x) = S_e \equiv \text{const} \end{cases}. \tag{8.12}$$

The flux-based wave decomposition method is well-balanced if the left and right source term approximations $\vec{\Psi}_{i\pm 1/2}$ are chosen as follows:

$$\vec{\Psi}_{i-1/2} = \begin{pmatrix} 0 \\ -g \cdot \dfrac{h_i + h_{i-1}}{2} \cdot \dfrac{B_i - B_{i-1}}{\Delta x} \end{pmatrix} \tag{8.13}$$

and

$$\vec{\Psi}_{i+1/2} = \begin{pmatrix} 0 \\ -g \cdot \dfrac{h_{i+1} + h_i}{2} \cdot \dfrac{B_{i+1} - B_i}{\Delta x} \end{pmatrix}. \tag{8.14}$$

It can be verified by direct substitution that when (8.12) takes place and source terms are chosen as (8.13) and (8.14), then both (8.5) and (8.6) are satisfied exactly.

According to [3], the flux-based wave decomposition approach for the shallow water equations is extensively used in tsunami simulation, an application where it is particularly critical that small perturbations around the ocean-at-rest state are accurately captured since the magnitude of a tsunami wave is generally one meter or even less while the bathymetry varies on the order of several kilometers.

9 Implementation

This paper is supplemented with an open-source implementation which can be found at the online Git repository (**https://bitbucket.org/Alexander-Shukaev/shallow-water-equations**). To obtain the source code, one can either follow this link and download the archive, or one can directly use Git to clone the contents of repository:

```
git clone https://bitbucket.org/Alexander-Shukaev/shallow-water-equations.git
```

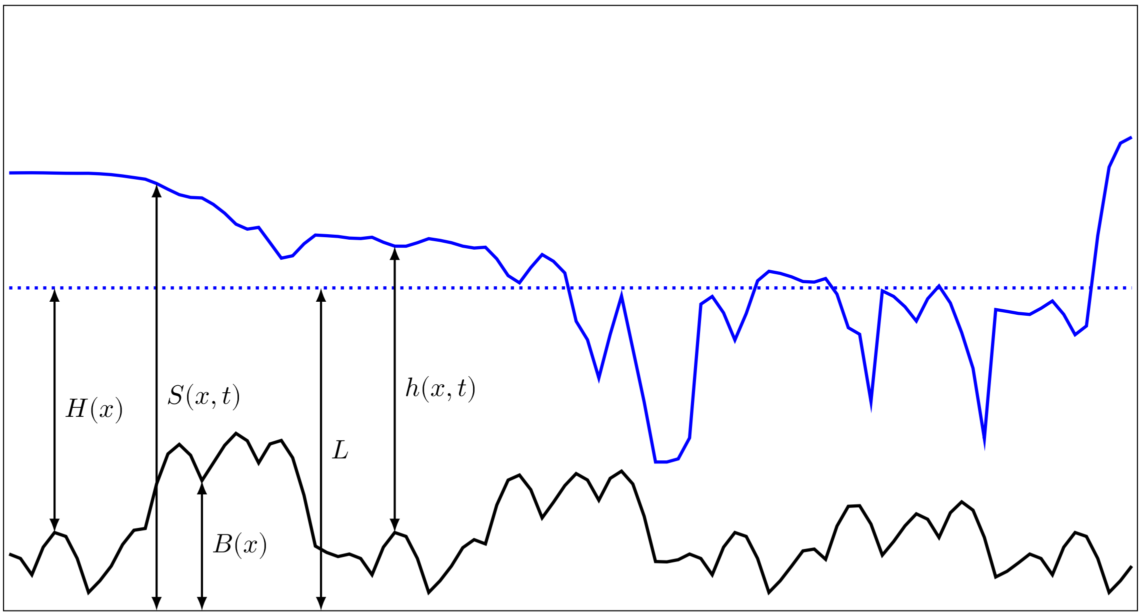

Figure 8.1 Shallow water equations with bathymetry

The implementation consists of flux-based wave decomposition algorithm (see **Section 7**) with arbitrarily complex bathymetry support based on well-balanced discretization approach (see **Section 8**). The important quantities are plotted in the real-time as the time marching goes. Furthermore, there are plenty of scenarios supplied out-of-the-box (see **Figure 9.1**). However, thanks to generic interface for scenario construction, it is straightforward to assemble a custom scenario.

The merit of this implementation is that the code is kept as simple as possible as the goal was to maintain readability and close conformance to formulas. In other words, its purpose is purely educational. Here is the summary of key points:

- implemented with a simple programming language, namely MATLAB;
- strives to stick to the formulas presented in this paper as close as possible, in particular:
  - Godunov's method is implemented according to **(7.10)**;
  - left and right averaged Jacobians are chosen to be **(7.11)** and **(7.12)** respectively;
  - flux waves are computed using **(8.3)** and **(8.4)**;
  - left and right source term approximations are treated as shown in **(8.13)** and **(8.14)** respectively.

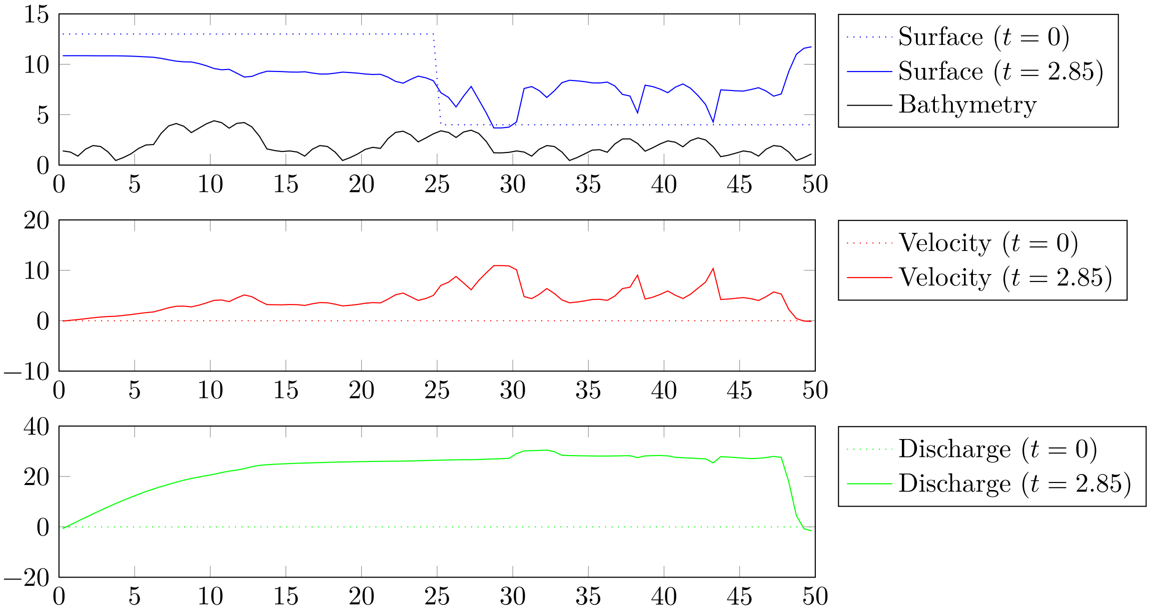

Figure 9.1 An example of scenario: dam break over arbitrarily complex bathymetry with walls on the boundaries

- not cluttered with a single optimization to preserve readability;
- the code works much slower than it could, but the real-time animation with acceptable frame rate is still achieved with moderate number of cells;
- dry interfaces are not handled as this adds some nontrivial complexity to the code; therefore one has to carefully design the scenario, so that drying of cells does not happen during simulation.

10 Results

In case of homogeneous (without bathymetry) shallow water equations **(6.4)**, all the tested scenarios (see **Section 9**) show results that are consistent with the theory presented in this paper. For instance, the classical *dam break* scenarios (refer to [1] for details) described in files `scenario1.m` (with outflow on the boundaries) and `scenario2.m` (with velocity reflection on the boundaries, i.e. solid walls) behave as expected: producing one propagating shock wave and one evolving rarefaction wave as the time marching goes. The two-sided dam break problem (`scenario3.m`) is consistent as well, yielding two shock waves and two rarefaction waves respectively.

Introducing several dam break problems simultaneously (`scenario4.m`) results in more complex wave interaction. However, one can notice that at any moment in time the solution still consists only of the two types of waves, namely shock waves and rarefaction waves, which is once again consistent with the theory (see [1]).

As for the inhomogeneous (with bathymetry) shallow water equations **(8.10)**, all the tested scenarios have shown consistent results too. For example, the dam break problem over the arbitrary bathymetry (`scenario7.m`), which is depicted in **Figure 9.1**, demonstrates how shock wave evolves as soon as some wave passes a hump, and the strength of the shock wave evolved depends on the height of the hump.

Finally, there is an extremely important test case described in `scenario5.m` to determine whether the flux-based wave decomposition approach is well-balanced as promised by theory (see **Section 8**). As expected, the initial surface-at-rest state is preserved proving that the numerical method developed in this paper is indeed well-balanced with the proper choice of source term averages (see **Section 8**). One interesting observation though, is that the velocity still has infinitesimal perturbations whose order of magnitude is exactly of the one of the floating-point precision being used during the simulation. Although theoretically left-hand sides of **(8.3)** and **(8.4)** are supposed to exactly cancel out in case of equilibrium state (see **Section 8**), and they indeed do in exact arithmetics, in general they of course cannot cancel out exactly in finite precision arithmetics (such as floating-point arithmetics). However, we still obtain good results since the fluid surface stays in rest anyway, and the velocity perturbations of that magnitude are negligible for real-life applications.

References

[1] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems.* (Cambridge University Press, 2002).

[2] D. Bale, R. J. LeVeque, S. Mitran, and J. A. Rossmanith, *A wave propagation method for conservation laws and balance laws with spatially varying flux functions*, SIAM Journal on Scientific Computing **24** (2002), 955-978

[3] R. J. LeVeque, *A well-balanced path-integral f-wave method for hyperbolic problems with source terms*, Journal of Scientific Computing **48** (2010), 209-226

[4] R. J. LeVeque, *Balancing source terms and flux gradients in high-resolution godunov methods: The quasi-steady wave-propagation algorithm*, Journal of Computational Physics **146** (1998), 346-365

[5] T. Hou and P. LeFloch, *Why nonconservative schemes converge to the wrong solutions: Error analysis*, Mathematics of Computation **62** (1994), 497-530