# Adaptive linearization for the optimal control problem of gas flow in pipeline networks

Oliver Kolb, Jens Lang, Pia Bales

TECHNISCHE UNIVERSITÄT DARMSTADT

## Abstract

We are interested in simulation and optimization of gas networks. Usually, a gas network consists of various components like compressors and valves connected by pipes. The aim is to run the network cost efficiently whereas the demands of consumers have to be satisfied. This results in a complex nonlinear mixed integer problem. We address this task with methods provided by discrete optimization. Therefore, the gas dynamics in all pipes and at compressors must be described by piecewise linear constraints. We introduce an adaptive approach for the linearization process to handle the complexity on the one hand and the aimed accuracy on the other. Further, we present numerical simulation and optimization results based on our model.


## 1 Introduction

There has been intense research in the field of simulation and optimization of gas transport in networks during the last years. Especially for the optimization task many approaches neglect different aspects - e.g. the stationary case is considered as in [11] or binary decisions like switching processes for compressor stations must be determined before the optimization process. We are interested in the transient case of gas network optimization and aim to treat continuous as well as integer variables in our problems. Therefore, we apply methods provided by discrete optimization as described in [12].

Discrete optimization deals with mixed integer linear problems (MILPs) and has many applications in the field of operations research. Among these are optimization tasks for scheduling and transport processes like time tables for airports and train stations. In applications of practical relevance this results in large systems with thousands of variables. The basic tool of discrete optimization is the simplex algorithm. Integer variables are treated using relaxation as well as branch and cut techniques. Besides the integration of integer variables, the guaranty of global optimality is another main advantage of the algorithms used in discrete optimization.

But applying discrete optimization techniques to the transient case of gas network optimization yields two major problems: While only linear constraints can be posed in a linear mixed integer approach, components like compressor stations are described by nonlinear equations and moreover, one has to cope with the underlying PDEs of gas dynamics (see section 3). In general, nonlinearities in discrete optimization can be addressed with piecewise linearization, introducing approximation errors, new variables for every grid point in the linearized function and lots of equations describing the so-called SOS constraint. Thus, especially multidimensional nonlinear functions demand a trade-off between approximation accuracy and complexity of the MILP.

At first, we need a discretization of the underlying hyperbolic PDEs which allows for a linearization with reasonable effort (see section 4). We have examined a fully implicit box scheme and it proved to be a reliable basis for our optimization framework. Stability and convergence results of our scheme are presented in section 5. The next crucial step is the linearization process for the discretized PDE and other nonlinear constraints. We introduce an adaptive approach with independent error estimators for both the discretization and linearization errors (see section 6 and 7). For testing purposes, we have implemented our model in the form of a black-box simulator within an approved optimization framework [13]. Numerical results are presented in section 8.

## 2 Task definition

The aim of gas network optimization is to run a network cost efficiently whereas demands of consumers have to be satisfied. Here, the costs consist of the entire fuel gas consumption of all compressor stations $c \in E_C$:

$$\text{fuel gas} = \sum_{c \in E_C} \int_{t_{begin}}^{t_{end}} F_c(t)\,dt$$

where $F_c(t)$ denotes the fuel gas consumption of compressor $c$ at time $t$.

The consumer demands are described by time-dependent flux and target pressure values at all sinks $s \in V_S$. For a given compressor control the violation of the target pressure values is measured as follows:[1]

$$\text{gap to target} = \sum_{s \in V_S} \int_{t_{begin}}^{t_{end}} \left(p_{s,target}(t) - p_s(t)\right)_+ dt$$

where $p_s(t)$ and $p_{s,target}(t)$ denote the (target) pressure at sink $s$ at time $t$.

## 3 Model

We model a gas network as a directed, finite graph $G = (V, E)$. The edges $E$ correspond to the different components of the network and the vertices $V$ are inner/coupling or boundary nodes. For each node $v \in V$, $\delta_v^+$ denotes the set of outgoing edges and $\delta_v^-$ the set of ingoing edges. For the edges $e \in E$, we define disjoint intervals $[x_e^a, x_e^b]$, where $x_e^a$ belongs to the beginning of the edge $e$ and $x_e^b$ to the end.

The gas dynamics inside the pipes $E_p \subseteq E$ are described by the isothermal Euler equations with a friction term:

$$\partial_t (\rho A) + \partial_x (\rho_0 q) = 0 \tag{1}$$

$$\partial_t (\rho_0 q) + \partial_x \left( Ap + \frac{(\rho_0 q)^2}{\rho A} \right) = -\lambda(q) \frac{\rho_0 q |\rho_0 q|}{2 d \rho A} \tag{2}$$

where $p$ denotes the pressure, $\rho$ the density, $\rho_0$ the standard density of the gas, $q$ the flux, $\lambda$ the friction coefficient, $A$ the cross-sectional area and $d$ the diameter of the pipe. The friction factor is given by the implicit formula of Colebrook:

$$\frac{1}{\sqrt{\lambda}} = -2 \log_{10} \left( \frac{2.51}{Re \sqrt{\lambda}} + \frac{k}{3.71 d} \right) \tag{3}$$

$$Re = \frac{4 \rho_0}{\pi \eta d} \cdot |q| \tag{4}$$

where $\eta$ is a gas dependent constant (dynamic viscosity) and $k$ describes the roughness of the pipe. As equation of state we use

$$p = c^2 \rho \tag{5}$$

with a constant $c$, whereas our implementation also allows a nonlinear equation of state as in [12]. At coupling nodes of the network, we have to pose coupling conditions for all incoming and outgoing edges. According to [2] and [12], we claim equality of pressure as well as conservation of mass for all inner nodes $v \in V_{\text{inner}}$ and time $t \in [t_{\text{begin}}, t_{\text{end}}]$:

$$\sum_{i \in \delta_v^-} q(x_i^b, t) = \sum_{j \in \delta_v^+} q(x_j^a, t) \tag{6}$$

$$\forall i \in \delta_v^- \text{ and } \forall j \in \delta_v^+ : \quad p(x_i^b, t) = p(x_j^a, t) \quad . \tag{7}$$

For each compressor station $c \in E_C$, we apply the following two constraints for the compressor power and the fuel gas consumption (compare [6]):[2]

$$P(p_{in}, p_{out}, q_{in}) = d_P \cdot q_{in} \cdot \left( \left( \frac{p_{out}}{p_{in}} \right)^{\frac{\gamma - 1}{\gamma}} - 1 \right) \stackrel{!}{=} N \tag{8}$$

$$q_{out} = q_{in} - F(p_{in}, p_{out}, q_{in}) \tag{9}$$

$$F(p_{in}, p_{out}, q_{in}) = d_F \cdot q_{in} \cdot \left( \left( \frac{p_{out}}{p_{in}} \right)^{\frac{\gamma - 1}{\gamma}} - 1 \right) \tag{10}$$

[1] $(x)_+ = max(x, 0)$ .

[2] $d_P$ and $d_F$ may depend on the pressure if a nonlinear equation of state is used instead of (5).

where $p_{in}(t) = p(x_c^a, t)$, $p_{out}(t) = p(x_c^b, t)$, $q_{in}(t) = q(x_c^a, t)$ and $q_{out}(t) = q(x_c^b, t)$. $d_P$, $d_F$ and $\gamma$ are compressor and gas dependent constants and $N = N(t)$ denotes the current compressor power, a time dependent control variable.

## 4 Discretization

For the discretization of the isothermal Euler equations (1) and (2), we use a fully implicit box scheme which is symmetric in space. The general case is treated in the next section. Applied to the isothermal Euler equations, our scheme reads as follows:

$$\begin{aligned} \frac{\rho(x,t+\Delta t)+\rho(x+\Delta x,t+\Delta t)}{2} &= \frac{\rho(x,t)+\rho(x+\Delta x,t)}{2} && (11)\\ &\quad -\frac{\Delta t}{\Delta x}\left(f_1(x+\Delta x,t+\Delta t)-f_1(x,t+\Delta t)\right) \\ \frac{q(x,t+\Delta t)+q(x+\Delta x,t+\Delta t)}{2} &= \frac{q(x,t)+q(x+\Delta x,t)}{2} && (12)\\ &\quad -\frac{\Delta t}{\Delta x}\left(f_2(x+\Delta x,t+\Delta t)-f_2(x,t+\Delta t)\right) \\ &\quad -\Delta t\left(\frac{fric(x,t+\Delta t)+fric(x+\Delta x,t+\Delta t)}{2}\right) \end{aligned}$$

where

$$\begin{aligned} f_1(x,t) &= \frac{\rho_0}{A}q(x,t) \\ f_2(x,t) &= \frac{1}{\rho_0}\left(Ap(x,t)+\frac{(\rho_0 q(x,t))^2}{\rho(x,t)A}\right) \\ fric(x,t) &= \lambda\frac{\rho_0 q(x,t)|q(x,t)|}{2d\rho(x,t)A}\,. \end{aligned}$$

The resulting set of implicit equations is solved together with other constraints like coupling and boundary conditions at the nodes and the compressor equations (8) and (9). This is done using an adapted version of Newton's method and applying sparse matrix techniques [5].

## 5 Stability analysis and convergence results

For a general balance law of the form $u_t + f(u)_x = g(u)$ (with $f \in C^1$), our discretization scheme reads as follows:

$$\frac{u_{j-1}^{n+1}+u_j^{n+1}}{2} = \frac{u_{j-1}^{n}+u_j^{n}}{2} - \frac{\Delta t}{\Delta x}\left(f(u_j^{n+1})-f(u_{j-1}^{n+1})\right) + \Delta t\,\frac{g(u_{j-1}^{n+1})+g(u_j^{n+1})}{2}\,. \tag{13}$$

When implementing this method for a scalar balance law on a finite grid $[x_l, x_r]$, 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, respectively the sign of $f'$. Accordingly, the sign of $f'$ must not change in the interval and we will require such a constraint in the proofs later. In our case, the isothermal Euler equations, we are dealing with a system of balance laws. For systems of balance laws, the signature of the characteristic directions must not change. This applies to the practical relevant cases ($|v| \ll c$) since the eigenvalues of the Jacobian $f_u$ are $\lambda_{1/2} = v \pm c$.

The analysed properties of our fully implicit box scheme are motivated by the result of Kružkov given below for scalar balance laws on unbounded domains (see [9]). Therefore, we consider scalar balance laws of the form

$$u_t + f(u)_x = g(u), \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+ \tag{14}$$

with given initial data

$$u(x,0) = u_0(x), \quad x \in \mathbb{R}\,. \tag{15}$$

We are interested in weak solutions of the Cauchy problem (14)-(15) that is a function $u \in L^\infty(\mathbb{R}\times\mathbb{R}_+)$ satisfying:

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}[u\phi_t + f(u)\phi_x]\,dt\,dx + \int_{-\infty}^{\infty} u_0(x)\phi(x,0)\,dx = -\int_{-\infty}^{\infty}\int_{0}^{\infty} g(u)\phi\,dt\,dx \qquad \forall \phi \in C_0^1\left(\mathbb{R}\times\mathbb{R}_+\right)\ . \tag{16}$$

For given $\Delta t$ and $\Delta x$, we identify our discrete approximate solutions given by the scheme (13) with piecewise constant functions $\tilde{u}$ in the following way:

$$\tilde{u}(x,t) = u_j^n \quad \text{for } (x,t) \in I_j \times J_n \tag{17}$$

where $I_j = [(j-0.5)\Delta x, (j+0.5)\Delta x)$ and $J_n = [n\Delta t, (n+1)\Delta t)$. For the initial conditions, we set

$$u_j^0 = \int_{I_j} u_0(x)\,dx\ . \tag{18}$$

In general, the solution of (16) is not unique and the physical one is characterized by the following entropy condition:

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}[\eta(u)\phi_t + F(u)\phi_x]\,dt\,dx + \int_{-\infty}^{\infty} \eta(u_0(x))\phi(x,0)\,dx \geq -\int_{-\infty}^{\infty}\int_{0}^{\infty} \eta'(u)g(u)\phi\,dt\,dx \qquad \forall \phi \in C_0^1\left(\mathbb{R}\times\mathbb{R}_+\right), \phi \geq 0 \tag{19}$$

where $\eta \in C^2(\mathbb{R})$ is a strictly convex function and the entropy flux function $F$ satisfies $F'(u) = \eta'(u)f'(u)$ for all $u \in \mathbb{R}$.

**Theorem (Kružkov):** If $u_0 \in L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$, $f, g \in C^1(\mathbb{R})$, $g(0) = 0$ and $g' \leq 0$ holds, then the problem (16) possesses a unique entropy solution $u(x,t) = S(t)u_0$ satisfying

- $||S(t)u_0||_{L^\infty(\mathbb{R})} \leq ||u_0||_{L^\infty(\mathbb{R})}$
- $||S(t)u_0 - S(t)v_0||_{L^1(\mathbb{R})} \leq ||u_0 - v_0||_{L^1(\mathbb{R})} \quad \forall v_0 \in L^\infty(\mathbb{R})$
- $TV(S(t)u_0) \leq TV(u_0)$

## Existence of a unique solution

First of all, we show that our discretization scheme admits a unique solution in $L^1(\mathbb{Z})$ in every time step. As mentioned at the beginning, we will require that the sign of $f'$ does not change and therefore, we assume $f' \geq \lambda_{min} > 0$. The case $f' \leq -\lambda_{min} < 0$ can be treated analogously: While in the following proofs our scheme is always solved for $u_j^{n+1}$, one simply has to solve it for $u_{j-1}^{n+1}$.

**Proposition 1 (existence and uniqueness):** For $u^n \in L^1(\mathbb{Z})$, $g(0) = 0$, $g' \leq 0$, $f' \geq \lambda_{min} > 0$ and $\frac{\Delta t}{\Delta x} \geq \frac{1}{2\cdot\lambda_{min}}$ the scheme (13) admits a unique solution $u^{n+1} \in L^1(\mathbb{Z})$.

*Proof:* Algebraic transformations of our scheme (13) lead to:

$$\frac{u_{j-1}^{n+1} + u_j^{n+1}}{2} = \frac{u_{j-1}^n + u_j^n}{2} - \frac{\Delta t}{\Delta x}\left(f(u_j^{n+1}) - f(u_{j-1}^{n+1})\right) + \Delta t\,\frac{g(u_{j-1}^{n+1}) + g(u_j^{n+1})}{2}$$

$$\Leftrightarrow \quad u_j^{n+1} = \frac{1}{2}\left(u_{j-1}^n + u_j^n\right) - \frac{\Delta t}{\Delta x}\left(h(u_j^{n+1}) - h(u_{j-1}^{n+1})\right) + \Delta t\,\frac{g(u_{j-1}^{n+1}) + g(u_j^{n+1})}{2}$$

with $h(u) = f(u) - \frac{\Delta x}{2\Delta t}u$. Due to the requirements on $f'$ and $\frac{\Delta t}{\Delta x}$, we have $h' \geq 0$.

We introduce the following two operators $W, T : L^1(\mathbb{Z}) \to L^1(\mathbb{Z})$:

$$W(u)_j = \frac{1}{2}\left(u_{j-1} + u_j\right)$$

$$T(u)_j = u_j + \frac{\Delta t}{\Delta x}\left(h(u_j) - h(u_{j-1})\right) - \Delta t\,\frac{g(u_{j-1}^{n+1}) + g(u_j^{n+1})}{2}\ .$$

With these operators, our scheme (13) can be written as follows:

$$T(u^{n+1}) = W(u^n)\ .$$

To prove the existence of a unique solution, it suffices to show that $\|T'(u)\|_{L^1(\mathbb{Z})} \geq C > 0$ and that the linear operator $T'(u)$ is surjective. For an arbitrary $w \in L^1(\mathbb{Z})$, we have

$$\left(T'(u)w\right)_j = \underbrace{[1 + \underbrace{\frac{\Delta t}{\Delta x}h'(u_j)}_{\geq 0} - \underbrace{\frac{\Delta t}{2}g'(u_j)}_{\geq 0}]}_{=\alpha_j} w_j - \underbrace{[\underbrace{\frac{\Delta t}{\Delta x}h'(u_{j-1})}_{\geq 0} + \underbrace{\frac{\Delta t}{2}g'(u_{j-1})}_{\leq 0}]}_{=\beta_{j-1}} w_{j-1}$$

Since $|\beta_{j-1}| \leq \alpha_{j-1}$, applying the triangle inequality and summing up over $j \in \mathbb{Z}$ yields:

$$\begin{aligned}
& |\left(T'(u)w\right)_j| \geq (1+\alpha_j)|w_j| - \alpha_{j-1}|w_{j-1}| \\
\Rightarrow\quad & \|T'(u)w\|_{L^1(\mathbb{Z})} \geq \|w\|_{L^1(\mathbb{Z})} \\
\Rightarrow\quad & \|T'(u)\|_{L^1(\mathbb{Z})} \geq 1 =: C\ .
\end{aligned}$$

*Remark:* $(\alpha_j) \in L^\infty(\mathbb{Z})$ since $u \in L^1(\mathbb{Z}) \subset L^\infty(\mathbb{Z})$ and $h', g' \in C(\mathbb{R})$.

For the proof of the surjectivity of $T'(u)$, we consider an arbitrary $y \in L^1(\mathbb{Z})$ and the sequence $(w^k)_{k\in\mathbb{N}} \subset L^1(\mathbb{Z})$ defined as follows:

$$w_j^k = \begin{cases} 0 & \text{if } j \leq -k-1 \\ \dfrac{1}{1+\alpha_j}(y_j + \beta_{j-1}w_{j-1}) & \text{if } j \geq -k \end{cases}\ .$$

First, we show that $w^k \in L^1(\mathbb{Z})$. For $j \geq -k$, we have:

$$\begin{aligned}
& (1+\alpha_j)w_j^k = y_j + \beta_{j-1}w_{j-1} \\
\Rightarrow\quad & (1+\alpha_j)|w_j^k| \leq |y_j| + |\beta_{j-1}||w_{j-1}| \leq |y_j| + \alpha_{j-1}|w_{j-1}|\ .
\end{aligned}$$

Since $w_j^k = 0$ for $j \leq -k-1$, summation yields:

$$\begin{aligned}
& \sum_{j=-\infty}^{N}(1+\alpha_j)|w_j^k| \leq \sum_{j=-\infty}^{N}|y_j| + \sum_{j=-\infty}^{N}\alpha_{j-1}|w_{j-1}| \\
\Rightarrow\quad & \sum_{j=-\infty}^{N}|w_j^k| \leq \sum_{j=-\infty}^{N}|y_j| - \alpha_N|w_N^k| \leq \sum_{j=-\infty}^{N}|y_j| \qquad (20) \\
\Rightarrow\quad & \|w^k\|_{L^1(\mathbb{Z})} \leq \|y\|_{L^1(\mathbb{Z})}\ .
\end{aligned}$$

Next, we show that $(w^k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $L^1(\mathbb{Z})$. Let $m, n \in \mathbb{N}$, $m > n$. For $j \geq -n$, we have:

$$\begin{aligned}
& (1+\alpha_j)(w_j^m - w_j^n) = \beta_{j-1}(w_{j-1}^m - w_{j-1}^n) \\
\Rightarrow\quad & (1+\alpha_j)|w_j^m - w_j^n| = |\beta_{j-1}||w_{j-1}^m - w_{j-1}^n| \leq \alpha_{j-1}|w_{j-1}^m - w_{j-1}^n|\ .
\end{aligned}$$

With $\alpha_\infty := \|\alpha\|_{L^\infty(\mathbb{Z})}$, summation yields:

$$\begin{aligned}
& \sum_{j=-n}^{N}(1+\alpha_j)|w_j^m - w_j^n| \leq \sum_{j=-n}^{N}\alpha_{j-1}|w_{j-1}^m - w_{j-1}^n| \\
\Rightarrow\quad & \sum_{j=-n}^{N}|w_j^m - w_j^n| \leq \alpha_{-n-1}|w_{-n-1}^m - \underbrace{w_{-n-1}^n}_{=0}| - \alpha_N|w_N^m - w_N^n| \leq \alpha_\infty|w_{-n-1}^m| \\
\Rightarrow\quad & \sum_{j=-\infty}^{\infty}|w_j^m - w_j^n| = \sum_{j=-n}^{\infty}|w_j^m - w_j^n| + \sum_{j=-\infty}^{-n-1}|w_j^m - \underbrace{w_j^n}_{=0}| \leq \alpha_\infty|w_{-n-1}^m| + \sum_{j=-\infty}^{-n-1}|w_j^m| \\
\Rightarrow\quad & \|w^m - w^n\|_{L^1(\mathbb{Z})} = \sum_{j=-\infty}^{\infty}|w_j^m - w_j^n| \leq (1+\alpha_\infty)\sum_{j=-\infty}^{-n-1}|w_j^m| \overset{(20)}{\leq} (1+\alpha_\infty)\sum_{j=-\infty}^{-n-1}|y_j|
\end{aligned}$$

Since $y \in L^1(\mathbb{Z})$, we know that for all $\epsilon > 0$ there exists an $N(\epsilon) \in \mathbb{N}$ such that for all $i \geq N(\epsilon)$:

$$\sum_{j=-\infty}^{-i-1} |y_j| \leq \frac{1}{1+\alpha_\infty}\epsilon \ .$$

Therefore, $(w^k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $L^1(\mathbb{Z})$, because for all $m,n > N(\epsilon)$: $\|w^m - w^n\|_{L^1(\mathbb{Z})} \leq \epsilon$. Since $L^1(\mathbb{Z})$ is a Banach space, there exists $w^* = \lim_{k\to\infty} w^k$ (in $L^1(\mathbb{Z})$). Due to the continuity of $T'(u)$, we get:[3]

$$\|T'(u)w^* - y\|_{L^1(\mathbb{Z})} = \|T'(u)(\lim_{k\to\infty} w^k) - y\|_{L^1(\mathbb{Z})} = \lim_{k\to\infty} \|T'(u)w^k - y\|_{L^1(\mathbb{Z})} = \lim_{k\to\infty} \sum_{j=-\infty}^{-k-1} |y_j| = 0 \ .$$

Now, that we have shown that there exists a (unique) solution $u^{n+1} \in L^1(\mathbb{Z})$ for $u^n \in L^1(\mathbb{Z})$, it is reasonable to mention that our scheme is conservative for $g \equiv 0$:

$$\sum_{j=-\infty}^{\infty} u_j^{n+1} = \sum_{j=-\infty}^{\infty} \frac{u_{j-1}^{n+1} + u_j^{n+1}}{2} = \sum_{j=-\infty}^{\infty} \frac{u_{j-1}^{n} + u_j^{n}}{2} - \frac{\Delta t}{\Delta x} \underbrace{\sum_{j=-\infty}^{\infty} \left( f(u_j^{n+1}) - f(u_{j-1}^{n+1}) \right)}_{=0} = \sum_{j=-\infty}^{\infty} u_j^n$$

## Stability properties

For the following proofs, we introduce the operator

$$\left(T_\mu^h(u)\right)_j = u_j - \mu \frac{\Delta t}{\Delta x} \left( h(u_j) - h(u_{j-1}) \right)$$

with $\mu > 0$ and $h \in C^1$.

**Lemma 1:** For $u \in L^\infty(\mathbb{Z})$, $h \in C^1$ with $h' \geq 0$ and $\mu$ sufficiently small, we get the following stability properties for the operator $T_\mu^h$:

- $\|T_\mu^h(u)\|_{L^\infty(\mathbb{Z})} \leq \|u\|_{L^\infty(\mathbb{Z})}$

and if $u, v \in L^1(\mathbb{Z})$:

- $\|T_\mu^h(u) - T_\mu^h(v)\|_{L^1(\mathbb{Z})} \leq \|u - v\|_{L^1(\mathbb{Z})}$
- $TV(T_\mu^h(u)) \leq TV(u)$ .

*Proof:* First, we show that the function $\Phi(u_{j-1}, u_j, u_{j+1}) := \left(T_\mu^h(u)\right)_j$ is non-decreasing in every argument for $\mu = \frac{1}{h_\infty}$ with $h_\infty = \max(1, \frac{\Delta t}{\Delta x} \sup_{j\in\mathbb{Z}} h'(u_j))$:[4]

$$\begin{aligned}
\frac{\partial}{\partial_1}\Phi(u_{j-1}, u_j, u_{j+1}) &= \mu \frac{\Delta t}{\Delta x} \underbrace{h'(u_{j-1})}_{\geq 0} \geq 0 \\
\frac{\partial}{\partial_2}\Phi(u_{j-1}, u_j, u_{j+1}) &= 1 - \underbrace{\mu \frac{\Delta t}{\Delta x} h'(u_j)}_{\leq 1} \geq 0 \\
\frac{\partial}{\partial_3}\Phi(u_{j-1}, u_j, u_{j+1}) &= 0 \ .
\end{aligned}$$

[3] Obviously, $\|T'(u)\|_{L^1(\mathbb{Z})} \leq 1 + 2\alpha_\infty$.

[4] $\sup_{j\in\mathbb{Z}} h'(u_j) < \infty$ since $u \in L^\infty(\mathbb{Z})$ and $h' \in C$.

Thus, for $\mu$ given as above, the operator $T_\mu^h$ can be interpreted as a so-called monotone scheme in conservation form (with numerical flux function $\tilde{h}(u_l,u_r)=h(u_l)$) and applying the results from [4] yields the inequalities above.

**Proposition 2 (stability):** For $u^n, v^n \in L^\infty(\mathbb{Z}) \cap L^1(\mathbb{Z}) = L^1(\mathbb{Z})$, the scheme (13) has the following stability properties:

a) If $\frac{\Delta x}{\Delta t} \leq 2f'(\cdot) + \Delta x g'(\cdot)$:

- $\|u^{n+1}\|_{L^\infty(\mathbb{Z})} \leq \|u^n\|_{L^\infty(\mathbb{Z})}$ .

b) If $\frac{\Delta x}{\Delta t} \leq 2f'(\cdot)$:

- $\|u^{n+1} - v^{n+1}\|_{L^1(\mathbb{Z})} \leq \|u^n - v^n\|_{L^1(\mathbb{Z})}$
- $TV(u^{n+1}) \leq TV(u^n)$ .

*Proof:* a) Similar to the proof of proposition 1, algebraic transformations of our scheme (13) lead to:

$$\begin{aligned} \frac{u_{j-1}^{n+1} + u_j^{n+1}}{2} &= \frac{u_{j-1}^n + u_j^n}{2} - \frac{\Delta t}{\Delta x}\left(f(u_j^{n+1}) - f(u_{j-1}^{n+1})\right) + \Delta t \frac{g(u_{j-1}^{n+1}) + g(u_j^{n+1})}{2} \\ \Leftrightarrow \quad u_j^{n+1} &= \frac{1}{2}\left(u_{j-1}^n + u_j^n\right) - \frac{\Delta t}{\Delta x}\left(\hat{h}(u_j^{n+1}) - \hat{h}(u_{j-1}^{n+1})\right) + \Delta t\, g(u_j^{n+1}) \end{aligned}$$

with $\hat{h}(u) = f(u) - \frac{\Delta x}{2\Delta t}u + \frac{\Delta x}{2}g(u)$. Due to the assumption $\frac{\Delta x}{\Delta t} \leq 2f'(\cdot) + \Delta x g'(\cdot)$, we have $\hat{h}'(\cdot) \geq 0$. Using the operator $T_\mu^{\hat{h}}$, the scheme (13) can be written as follows:

$$u_j^{n+1} = \frac{\mu}{1+\mu}W(u^n) + \frac{1}{1+\mu}T_\mu^{\hat{h}}(u^{n+1}) + \Delta t \frac{\mu}{1+\mu}g(u_j^{n+1}) .$$

The *Intermediate Value Theorem* yields $g(u_j^{n+1}) = g'(\xi_j^{n+1})u_j^{n+1}$ and therefore:

$$\begin{aligned} (1 - \Delta t \frac{\mu}{1+\mu}\underbrace{g'(\xi_j^{n+1})}_{\leq 0})u_j^{n+1} &= \frac{\mu}{1+\mu}W(u^n) + \frac{1}{1+\mu}T_\mu^{\hat{h}}(u^{n+1}) \\ \Rightarrow \quad |u_j^{n+1}| &\leq |\frac{\mu}{1+\mu}W(u^n) + \frac{1}{1+\mu}T_\mu^{\hat{h}}(u^{n+1})| \end{aligned}$$

Applying the triangle inequality and taking the supremum over $j \in \mathbb{Z}$ yields:

$$\|u^{n+1}\|_{L^\infty(\mathbb{Z})} \leq \frac{\mu}{1+\mu}\|W(u^n)\|_{L^\infty(\mathbb{Z})} + \frac{1}{1+\mu}\|T_\mu^{\hat{h}}(u^{n+1})\|_{L^\infty(\mathbb{Z})} .$$

Since we may choose a sufficiently small $\mu$, we obtain from lemma 1:

$$\begin{aligned} &\|u^{n+1}\|_{L^\infty(\mathbb{Z})} \leq \frac{\mu}{1+\mu}\|W(u^n)\|_{L^\infty(\mathbb{Z})} + \frac{1}{1+\mu}\|u^{n+1}\|_{L^\infty(\mathbb{Z})} \\ \Rightarrow \quad &\frac{\mu}{1+\mu}\|u^{n+1}\|_{L^\infty(\mathbb{Z})} \leq \frac{\mu}{1+\mu}\|W(u^n)\|_{L^\infty(\mathbb{Z})} \\ \Rightarrow \quad &\|u^{n+1}\|_{L^\infty(\mathbb{Z})} \leq \|W(u^n)\|_{L^\infty(\mathbb{Z})} \leq \|u^n\|_{L^\infty(\mathbb{Z})} . \end{aligned}$$

b) For the proof of $L^1$- and $TV$-stability, we use the same form of our scheme (13) as in proposition 1:

$$u_j^{n+1} = \frac{1}{2}\left(u_{j-1}^n + u_j^n\right) - \frac{\Delta t}{\Delta x}\left(h(u_j^{n+1}) - h(u_{j-1}^{n+1})\right) + \Delta t \frac{g(u_{j-1}^{n+1}) + g(u_j^{n+1})}{2}$$

with $h(u) = f(u) - \frac{\Delta x}{2\Delta t}u$. Due to the requirements on $f'$ and $\frac{\Delta t}{\Delta x}$, we have $h'(\cdot) \geq 0$. Using the operator $T^h_\mu$, we have:

$$u^{n+1}_j - v^{n+1}_j =$$
$$= \frac{\mu}{1+\mu}\left(W(u^n)_j - W(v^n)_j\right) + \frac{1}{1+\mu}\left(T^h_\mu(u^{n+1})_j - T^h_\mu(v^{n+1})_j\right) + \frac{\Delta t}{2}\frac{\mu}{1+\mu}\left(g(u^{n+1}_{j-1}) + g(u^{n+1}_j) - g(v^{n+1}_{j-1}) - g(v^{n+1}_j)\right).$$

Again, the *Intermediate Value Theorem* yields $g(u^{n+1}_j) - g(v^{n+1}_j) = g'(\xi^{n+1}_j)(u^{n+1}_j - v^{n+1}_j)$ and with $\gamma_j = -\frac{\Delta t}{2}\frac{\mu}{1+\mu}g'(\xi^{n+1}_j) \geq 0$, we get:

$$(1+\gamma_j)(u^{n+1}_j - v^{n+1}_j) = \frac{\mu}{1+\mu}\left(W(u^n)_j - W(v^n)_j\right) + \frac{1}{1+\mu}\left(T^h_\mu(u^{n+1})_j - T^h_\mu(v^{n+1})_j\right) - \gamma_{j-1}(u^{n+1}_{j-1} - v^{n+1}_{j-1})$$
$$\Rightarrow\ (1+\gamma_j)|u^{n+1}_j - v^{n+1}_j| \leq \frac{\mu}{1+\mu}|W(u^n)_j - W(v^n)_j| + \frac{1}{1+\mu}|T^h_\mu(u^{n+1})_j - T^h_\mu(v^{n+1})_j| + \gamma_{j-1}|u^{n+1}_{j-1} - v^{n+1}_{j-1}|.$$

For a sufficiently small $\mu$, summing up over $j \in \mathbb{Z}$ and subtracting $\sum_{j=-\infty}^{\infty} \gamma_j |u^{n+1}_j - v^{n+1}_j|$ from both sides yields:

$$\|u^{n+1} - v^{n+1}\|_{L^1(\mathbb{Z})} \leq \frac{\mu}{1+\mu}\|W(u^n) - W(v^n)\|_{L^1(\mathbb{Z})} + \frac{1}{1+\mu}\|(T^h_\mu(u^{n+1}) - T_\mu(v^{n+1})\|_{L^1(\mathbb{Z})}$$
$$\leq \frac{\mu}{1+\mu}\|W(u^n) - W(v^n)\|_{L^1(\mathbb{Z})} + \frac{1}{1+\mu}\|u^{n+1} - v^{n+1}\|_{L^1(\mathbb{Z})}$$
$$\Rightarrow\ \|u^{n+1} - v^{n+1}\|_{L^1(\mathbb{Z})} \leq \|W(u^n) - W(v^n)\|_{L^1(\mathbb{Z})} \leq \|u^n - v^n\|_{L^1(\mathbb{Z})}.$$

Substituting $v^n_j = u^n_{j-1}$ ($\Rightarrow v^{n+1}_j = u^{n+1}_{j-1}$) immediately yields:

$$TV(u^{n+1}) \leq TV(u^n).$$

*Remark:* The assumptions for the proof of $L^1$- and $TV$-stability follow directly from the assumptions made in Proposition 1 about the existence and uniqueness of a solution in $L^1(\mathbb{Z})$. Therefore, the requirements for the proof of $L^\infty$-stability might be unexpected. Indeed, $\frac{\Delta t}{\Delta x} \geq \frac{1}{2\cdot\lambda_{min}}$ is also sufficient for $L^\infty$-stability if further assumptions are made, e.g. $u^{n+1} \geq 0$ or $u^{n+1} \leq 0$ (componentwise).

## Convergence results

In the following propositions, we will consider a sequence of approximate solutions $(u^{(k)})_{k\in\mathbb{N}}$ with mesh parameters $\Delta t^{(k)}$ and $\Delta x^{(k)}$, where $\Delta t^{(k)}, \Delta x^{(k)} \longrightarrow 0$ for $k \longrightarrow \infty$. Due to the definition (17), each $u^{(k)}$ can be interpreted as a function of $(x,t) \in \mathbb{R} \times \mathbb{R}_+$. For the sake of simplicity, we will leave out the parameter $k$ in the proofs.

In analogy to the *Lax-Wendroff-Theorem*, we state the following proposition:

**Proposition 3:** Let $(u^{(k)})_{k\in\mathbb{N}}$ be a sequence constructed by the scheme (13) and converging in $L^1_{\text{loc}}(\mathbb{R} \times \mathbb{R}_+)$ with $\Delta t^{(k)}, \Delta x^{(k)} \overset{k\to\infty}{\longrightarrow} 0$. Then, the limit $\hat{u} = \lim_{k\longrightarrow\infty} u^{(k)}$ is a weak solution of the Cauchy problem (14)-(15).

*Proof:* With $\phi \in C^1_0(\mathbb{R} \times \mathbb{R}_+)$ and $\phi^n_j = \phi(j\Delta x, n\Delta t)$, we obtain from (13):

$$\underbrace{\frac{u^{n+1}_{j-1} + u^{n+1}_j}{2}}_{=:\ u^{n+1}_{j-0.5}} = \underbrace{\frac{u^n_{j-1} + u^n_j}{2}}_{=:\ u^n_{j-0.5}} - \frac{\Delta t}{\Delta x}\Big(\underbrace{f(u^{n+1}_j)}_{=:\ f^{n+1}_j} - \underbrace{f(u^{n+1}_{j-1})}_{=:\ f^{n+1}_{j-1}}\Big) + \Delta t\,\underbrace{\frac{g(u^{n+1}_{j-1}) + g(u^{n+1}_j)}{2}}_{=:\ g^{n+1}_{j-0.5}}$$
$$\Rightarrow\ \Phi^{n+1}_{j-0.5}u^{n+1}_{j-0.5} = \Phi^{n+1}_{j-0.5}u^n_{j-0.5} - \frac{\Delta t}{\Delta x}\Phi^{n+1}_{j-0.5}\left(f^{n+1}_j - f^{n+1}_{j-1}\right) + \Delta t\,\Phi^{n+1}_{j-0.5}g^{n+1}_{j-0.5}.$$

Summation and (discrete) integration by parts yield:

$$\sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty}\Phi^{n+1}_{j-0.5}(u^{n+1}_{j-0.5} - u^n_{j-0.5}) + \frac{\Delta t}{\Delta x}\sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty}\Phi^{n+1}_{j-0.5}\left(f^{n+1}_j - f^{n+1}_{j-1}\right) = \Delta t\sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty}\Phi^{n+1}_{j-0.5}g^{n+1}_{j-0.5}$$
$$\Rightarrow\ \sum_{j=-\infty}^{j=\infty}\Phi^1_{j-0.5}u^0_{j-0.5} + \sum_{n=1}^{\infty}\sum_{j=-\infty}^{\infty}(\Phi^{n+1}_{j-0.5} - \Phi^n_{j-0.5})u^n_{j-0.5} + \frac{\Delta t}{\Delta x}\sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty}(\Phi^{n+1}_{j+0.5} - \Phi^{n+1}_{j-0.5})f^{n+1}_j = -\Delta t\sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty}\Phi^{n+1}_{j-0.5}g^{n+1}_{j-0.5}.$$

Multiplying with $\Delta x$ and taking the limit $k \longrightarrow \infty$, we finally get:

$$\Delta x \sum_{j=-\infty}^{j=\infty} \Phi^1_{j-0.5} u^0_{j-0.5} + \Delta x \Delta t \left[ \sum_{n=1}^{\infty} \sum_{j=-\infty}^{\infty} \frac{\Phi^{n+1}_{j-0.5} - \Phi^n_{j-0.5}}{\Delta t} u^n_{j-0.5} + \sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \frac{\Phi^{n+1}_{j+0.5} - \Phi^{n+1}_{j-0.5}}{\Delta x} f^{n+1}_j \right] = -\Delta x \Delta t \sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \Phi^{n+1}_{j-0.5} g^{n+1}_{j-0.5}$$

$$\stackrel{k\to\infty}{\longrightarrow} \int_{-\infty}^{\infty} \Phi(x,0) \underbrace{\hat{u}(x,0)}_{=u_0(x)} dx + \int_0^{\infty} \int_{-\infty}^{\infty} \left[ \Phi_t(x,t)\hat{u}(x,t) + \Phi_x(x,t) f(\hat{u}(x,t)) \right] dx\, dt = -\int_0^{\infty} \int_{-\infty}^{\infty} \phi(x,t) g(\hat{u}(x,t)) \, dx\, dt \, .$$

Next, we show convergence of our scheme to the entropy solution. For simplicity, we assume that

$$\frac{\Delta t}{\Delta x} = r = const.$$

As one can easily see from the proof, it is sufficient to claim $\frac{\Delta t}{\Delta x}, \frac{\Delta x}{\Delta t} \leq c < \infty$. Moreover, we claim the following for each $u \in (u^{(k)})_{k\in\mathbb{N}}$:

$$\left. \begin{array}{c} \|u^{n+1}\|_{L^\infty(\mathbb{Z})} \leq \|u^n\|_{L^\infty(\mathbb{Z})} \\ \|u^{n+1}\|_{L^1(\mathbb{Z})} \leq \|u^n\|_{L^1(\mathbb{Z})} \\ TV(u^{n+1}) \leq TV(u^n) \end{array} \right\} \tag{21}$$

This can be achieved by the assumptions made in proposition 2, but as mentioned in the remark, weaker conditions might be sufficient. Therefore, we rather claim these stability properties in proposition 4. Due to (18) and with the interpretation as piecewise constant functions (17), it follows immediately $\forall n \in \mathbb{N}$:[5]

$$\begin{array}{rcl} \|u^n\|_{L^\infty(\mathbb{R})} & \leq & \|u_0\|_{L^\infty(\mathbb{R})} \\ \|u^n\|_{L^1(\mathbb{R})} & \leq & \|u_0\|_{L^1(\mathbb{R})} \\ TV(u^n) & \leq & TV(u_0) \, . \end{array}$$

**Proposition 4:** Let $u_0 \in L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$, $f, g \in C^1(\mathbb{R})$, $g(0) = 0$, $g' \leq 0$, $f' \geq \lambda_{min} > 0$ and let $(u^{(k)})_{k\in\mathbb{N}}$ be a sequence constructed by the scheme (13) with initial conditions given by (18), fulfilling the stability properties (21) and $\Delta t^{(k)}, \Delta x^{(k)} \stackrel{k\to\infty}{\longrightarrow} 0$, $r = \frac{\Delta t^{(k)}}{\Delta x^{(k)}} \geq \frac{1}{2\cdot\lambda_{min}}$. Then, the limit $\hat{u} = \lim_{k\longrightarrow\infty} u^{(k)}$ exists (in $L^1_{\text{loc}}(\mathbb{R} \times \mathbb{R}_+)$) and $\hat{u}$ is the entropy solution of the Cauchy problem (14)-(15).

*Proof:* The proof is subdivided into the following parts: First, we show that there exists a convergent subsequence of our sequence in $L^1_{\text{loc}}(\mathbb{R} \times \mathbb{R}_+)$. Next, we proof that the limit $\hat{u}$ of every convergent subsequence fulfills the entropy condition (19). Then, together with the previous proposition, we have shown that our scheme converges to the entropy solution.

In the first part of the proof, we want to apply a compactness argument (compare [10]). Therefore, we consider the following function space:

$$L_{1,T} = \left\{ v : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R} \mid \|v\|_{1,T} < \infty \right\}$$
$$\text{with} \quad \|v\|_{1,T} = \int_0^T \int_{-\infty}^{\infty} |v(x,t)| \, dx\, dt \qquad .$$

It can be shown that the set

$$K = \left\{ v \in L_{1,T} : \; TV_T(v) \leq R \; \text{ and } \; \text{Supp}(v(\cdot,t)) \subseteq [-M,M] \;\; \forall t \in [0,T] \right\}$$
$$\text{with} \quad TV_T(v) = \limsup_{\epsilon\to 0} \frac{1}{\epsilon} \int_0^T \int_{-\infty}^{\infty} |v(x+\epsilon,t) - v(x,t)| \, dx\, dt + \limsup_{\epsilon\to 0} \frac{1}{\epsilon} \int_0^T \int_{-\infty}^{\infty} |v(x,t+\epsilon) - v(x,t)| \, dx\, dt$$

[5] Remark: For the total variation $TV$ the discrete definition given by $TV(v) = \sum_{j=-\infty}^{\infty} |v_j|$ is equal to the usual continuous definition.

is compact in $L_{1,T}$. For each of our piecewise constant approximations $u \in (u^{(k)})_{k\in\mathbb{N}}$, the following inequality holds:

$$\begin{aligned} TV_T(u) &\le \sum_{n=0}^{\lfloor T/\Delta t\rfloor}\sum_{j=-\infty}^{\infty}\left[\Delta t|u_{j+1}^n-u_j^n|+\Delta x|u_j^{n+1}-u_j^n|\right] \\ &\le \Delta t\sum_{n=0}^{\lfloor T/\Delta t\rfloor} TV(u^n)+\Delta x\sum_{n=0}^{\lfloor T/\Delta t\rfloor}\sum_{j=-\infty}^{\infty}|u_j^{n+1}-u_j^n| \\ &\le (T+\Delta t)\underbrace{TV(u^0)}_{\le TV(u_0)}+\Delta x\sum_{n=0}^{\lfloor T/\Delta t\rfloor}\sum_{j=-\infty}^{\infty}|u_j^{n+1}-u_j^n|\,. \end{aligned} \tag{22}$$

With

$$g'_\infty = \sup_{|w|\le\|u_0\|_{L^\infty(\mathbb{R})}}|g(w)| \quad\text{and}\quad f'_\infty = \sup_{|w|\le\|u_0\|_{L^\infty(\mathbb{R})}}|f'(w)|\,,$$

our scheme (13) yields:

$$\begin{aligned} & u_j^{n+1}-u_j^n = \frac12(u_j^{n+1}-u_{j-1}^{n+1})+\frac12(u_{j-1}^n-u_j^n)-\frac{\Delta t}{\Delta x}(f(u_j^{n+1})-f(u_{j-1}^{n+1}))+\frac{\Delta t}{2}(g(u_j^{n+1})+g(u_{j-1}^{n+1})) \\ \Rightarrow\ & |u_j^{n+1}-u_j^n|\le\frac12|u_j^{n+1}-u_{j-1}^{n+1}|+\frac12|u_{j-1}^n-u_j^n|+\frac{\Delta t}{\Delta x}\underbrace{|f(u_j^{n+1})-f(u_{j-1}^{n+1})|}_{\le f'_\infty|u_j^{n+1}-u_{j-1}^{n+1}|}+\frac{\Delta t}{2}(\underbrace{|g(u_j^{n+1})|}_{\le g'_\infty|u_j^{n+1}|}+\underbrace{|g(u_{j-1}^{n+1})|}_{\le g'_\infty|u_{j-1}^{n+1}|}) \\ \Rightarrow\ & \sum_{j=-\infty}^{\infty}|u_j^{n+1}-u_j^n|\le\frac12 TV(u^{n+1})+\frac12 TV(u^n)+\frac{\Delta t}{\Delta x}f'_\infty TV(u^{n+1})+\Delta t\, g'_\infty\|u^{n+1}\|_{L^1(\mathbb{Z})} \\ \Rightarrow\ & \sum_{j=-\infty}^{\infty}|u_j^{n+1}-u_j^n|\le(1+rf'_\infty)\underbrace{TV(u^0)}_{\le TV(u_0)}+rg'_\infty\underbrace{\Delta x\|u^0\|_{L^1(\mathbb{Z})}}_{\le\|u_0\|_{L^1(\mathbb{R})}}=:C \end{aligned}$$

Together with (22), we get:

$$TV_T(u) \le (T+\Delta t)TV(u_0)+\Delta x\sum_{n=0}^{\lfloor T/\Delta t\rfloor}C\le(T+\Delta t)\left(TV(u_0)+\frac1r C\right)=:R$$

Thus, if we restrict the functions $u^{(k)}$ to the domain $[-M,M]\times[0,T]$, they are contained in the compact set $K$. In other words, we can find a convergent subsequence in $L^1_{\mathrm{loc}}(\mathbb{R}\times\mathbb{R}_+)$.

From now on, we consider a convergent subsequence of our original sequence with limit $\hat{u}$. With $h(u)=f(u)-\frac{\Delta x}{\Delta t}u$, our scheme (13) reads:

$$\begin{aligned} & u_j^{n+1}-u_j^n+\frac{\Delta t}{\Delta x}(h(u_j^{n+1})-h(u_{j-1}^{n+1}))+\frac12(u_j^n-u_{j-1}^n)=\Delta t\frac{g(u_j^{n+1})+g(u_{j-1}^{n+1})}{2} \\ \Rightarrow\ & \eta'(u_j^{n+1})(u_j^{n+1}-u_j^n)+\frac{\Delta t}{\Delta x}\eta'(u_j^{n+1})(h(u_j^{n+1})-h(u_{j-1}^{n+1}))+\frac12\eta'(u_j^{n+1})(u_j^n-u_{j-1}^n)=\Delta t\,\eta'(u_j^{n+1})\frac{g(u_j^{n+1})+g(u_{j-1}^{n+1})}{2}\,. \end{aligned}$$

From the convexity of $\eta$ and the assumptions on $\frac{\Delta x}{\Delta t}$, we have:

$$\eta(u_j^{n+1})-\eta(u_j^n)\le\eta'(u_j^{n+1})(u_j^{n+1}-u_j^n)\quad\text{and}\quad\int_{u_{j-1}^{n+1}}^{u_j^{n+1}}\eta'(w)h'(w)\,dw\le\eta'(u_j^{n+1})(h(u_j^{n+1})-h(u_{j-1}^{n+1})) \tag{23}$$

and therefore:

$$\eta(u_j^{n+1})-\eta(u_j^n)+\frac{\Delta t}{\Delta x}\int_{u_{j-1}^{n+1}}^{u_j^{n+1}}\eta'(w)h'(w)\,dw+\frac12\eta'(u_j^{n+1})(u_j^n-u_{j-1}^n)\le\Delta t\,\eta'(u_j^{n+1})\frac{g(u_j^{n+1})+g(u_{j-1}^{n+1})}{2}\,.$$

By definition

$$h'(w) = f'(w) - \frac{\Delta x}{2\Delta t} \quad \text{and} \quad F'(w) = \eta'(w) f'(w)$$

and by applying the first inequality of (23) a second time, this yields

$$\underbrace{\eta(u_j^{n+1})}_{=:\eta_j^{n+1}} - \underbrace{\eta(u_j^n)}_{=:\eta_j^n} + \frac{\Delta t}{\Delta x}(\underbrace{F(u_j^{n+1})}_{=:F_j^{n+1}} - \underbrace{F(u_{j-1}^{n+1})}_{=:F_{j-1}^{n+1}}) + \underbrace{R(u_{j-1}^n, u_j^n, u_{j-1}^{n+1}, u_j^{n+1})}_{=:R_j^{n+1}} \leq \Delta t \eta'(u_j^{n+1}) \underbrace{\frac{g(u_j^{n+1}) + g(u_{j-1}^{n+1})}{2}}_{=:g_{j-0.5}^{n+1}}$$

where

$$R(u_{j-1}^n, u_j^n, u_{j-1}^{n+1}, u_j^{n+1}) \quad := \quad \frac{1}{2}\eta'(u_j^{n+1})(u_j^n - u_{j-1}^n - (u_j^{n+1} - u_{j-1}^{n+1})) .$$

With $\phi \in C_0^1(\mathbb{R} \times \mathbb{R}_+)$, $\phi \geq 0$ and $\phi_j^n = \phi(j\Delta x, n\Delta t)$, we get:

$$\sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \Phi_j^{n+1} \left(\eta_j^{n+1} - \eta_j^n\right) + \frac{\Delta t}{\Delta x} \sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \Phi_j^{n+1} \left(F_j^{n+1} - F_{j-1}^{n+1}\right) + \sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \Phi_j^{n+1} R_j^{n+1} \leq \Delta t \sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \Phi_j^{n+1} \eta'(u_j^{n+1}) g_{j-0.5}^{n+1} .$$

Similar to the proof of proposition 3, multiplication with $\Delta x$ and taking the limit $k \longrightarrow \infty$ yields:

$$\int_{-\infty}^{\infty} \int_0^{\infty} [\eta(\hat{u})\phi_t + F(\hat{u})\phi_x] \, dt \, dx + \int_{-\infty}^{\infty} \eta(u_0(x))\phi(x,0) \, dx \geq - \int_{-\infty}^{\infty} \int_0^{\infty} \eta'(\hat{u}) g(\hat{u}) \phi \, dt \, dx .$$

*Remark:* The "residual term" $\Delta x \sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \Phi_j^{n+1} R_j^{n+1}$ vanishes in the limit due to the bounded variation property of the sequence $u^{(k)}$ in time and space.

## 6 Linearization

Since we address the optimization task with methods provided by discrete optimization, all nonlinear terms in the discretized PDE as well as the other constraints must be approximated by piecewise linear functions. We will briefly explain the procedure for the two-dimensional case. The three-dimensional case, as needed for the compressor equations, and the one-dimensional case, which is only needed if a nonlinear equation of state is used, are similar.

The nonlinear terms $f_2$ and *fric* both depend on the pressure $p$ and flux $q$. So we need a triangulation of the feasible domain in the p-q-space. Assuming upper and lower bounds for both variables, a coarse uniform triangulation as shown in figure 1 can be used as initial grid. The nonlinear functions are evaluated at every grid point $(p_i, q_i)$. Finally, these are replaced by an affine combination of the precomputed values. For a nonlinear function $g(p,q)$ this reads as follows:

$$g(p,q) = g\left(\sum_{i\in\Lambda} \lambda_i p_i, \sum_{i\in\Lambda} \lambda_i q_i\right) \approx \sum_{i\in\Lambda} \lambda_i g(p_i, q_i) \quad \text{with} \quad \sum_{i\in\Lambda} \lambda_i = 1 \;\; \text{and} \;\; \lambda_i \geq 0 \;\; .$$

Moreover, all non-zero $\lambda_i$ must belong to a single triangle (SOS constraint).

Remark: Since the pressure and flux variables are space- and time-dependent a separate triangulation has to be stored for every grid point in the space-time-discretization.

## 7 Error estimators and refinement strategy

When we apply the piecewise linearized equations as constraints in an optimization framework, refinement is necessary to obtain the aimed accuracy. In general, a global refinement strategy is too expensive for larger problems. Therefore, we use local error estimators to find the right areas for refinement.

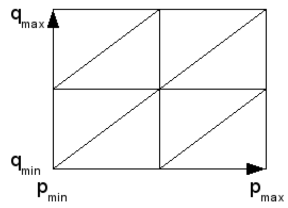


Figure 1: 2D-Triangulation.

In the presented framework, we have to distinguish between two kinds of errors: Discretization errors, which also occur in the nonlinear model, and linearization errors. For the refinement of the triangulations, it is important to have independent estimators for both: In case the linearization error becomes much smaller than the local truncation error due to the discretization, a further refinement of the triangulations only results in more computational effort but not more accuracy.

For a general balance law of the form $u_t + f(u)_x = g(u)$ our error estimator for the local discretization error reads as follows:

$$\begin{aligned} E^{j,n} = \frac{1}{12} \cdot \Big[ \; & u_{j+1}^{n+1} - u_{j+1}^{n-1} + 4 \cdot (u_j^{n+1} - u_j^{n-1}) + u_{j-1}^{n+1} - u_{j-1}^{n-1} \\ & + \frac{\Delta t}{\Delta x} \cdot \Big( f(u_{j+1}^{n+1}) - f(u_{j-1}^{n+1}) + 4 \cdot \big( f(u_{j+1}^{n}) - f(u_{j-1}^{n}) \big) + f(u_{j+1}^{n-1}) - f(u_{j-1}^{n-1}) \Big) \\ & - \frac{\Delta t}{3} \cdot \Big( g(u_{j-1}^{n-1}) + 4 \cdot g(u_j^{n-1}) + g(u_{j+1}^{n-1}) \Big) \\ & - 4 \cdot \frac{\Delta t}{3} \cdot \Big( g(u_{j-1}^{n}) + 4 \cdot g(u_j^{n}) + g(u_{j+1}^{n}) \Big) \\ & - \frac{\Delta t}{3} \cdot \Big( g(u_{j-1}^{n+1}) + 4 \cdot g(u_j^{n+1}) + g(u_{j+1}^{n+1}) \Big) \; \Big] . \end{aligned} \tag{24}$$

where $u_j^n = u(x_j, t_n)$. This error estimator is based on a weak local truncation error estimator presented in [7] and [8] where more details can be found for the conservative case ($g \equiv 0$). To estimate a local discretization error which is independent of linearization errors, we replace all nonlinear terms in (24) by the linearized ones.

The compressor equations (8) and (9) yield the following error estimators for the linearization error at a compressor:[6]

$$\begin{aligned} E_{p_{in}}^{c,n} &= \frac{P_{lin}(p_{in}^{c,n}, p_{out}^{c,n}, q_{in}^{c,n}) - P(p_{in}^{c,n}, p_{out}^{c,n}, q_{in}^{c,n})}{\frac{\partial}{\partial p_{in}} P(p_{in}^{c,n}, p_{out}^{c,n}, q_{in}^{c,n})} \\ E_{p_{out}}^{c,n} &= \frac{P_{lin}(p_{in}^{c,n}, p_{out}^{c,n}, q_{in}^{c,n}) - P(p_{in}^{c,n}, p_{out}^{c,n}, q_{in}^{c,n})}{\frac{\partial}{\partial p_{out}} P(p_{in}^{c,n}, p_{out}^{c,n}, q_{in}^{c,n})} \\ E_{q_{in}}^{c,n} &= \frac{P_{lin}(p_{in}^{c,n}, p_{out}^{c,n}, q_{in}^{c,n}) - P(p_{in}^{c,n}, p_{out}^{c,n}, q_{in}^{c,n})}{\frac{\partial}{\partial q_{in}} P(p_{in}^{c,n}, p_{out}^{c,n}, q_{in}^{c,n})} \\ E_{q_{out}}^{c,n} &= F_{lin}(p_{in}^{c,n}, p_{out}^{c,n}, q_{in}^{c,n}) - F(p_{in}^{c,n}, p_{out}^{c,n}, q_{in}^{c,n}) \end{aligned}$$

where the subindex "lin" denotes the linearized functions and $p_{in}^{c,n}$, $p_{out}^{c,n}$ and $q_{in}^{c,n}$ are the pressure and flux values at time $t_n$ and compressor number $c$. For the refinement decision, the absolute values of $E_{p_{in}}^{c,n}$ and $E_{q_{in}}^{c,n}$ are compared to the corresponding discretization errors at the end of all ingoing pipes as well as $E_{p_{out}}^{c,n}$ and $E_{q_{out}}^{c,n}$ are compared to the corresponding discretization errors at the beginning of all outgoing pipes.[7]

To estimate the linearization error (of a single triangulation) in the discretized momentum equation (12), we use:

$$E_q^{j,n} = \frac{\Delta t}{\Delta x} \cdot \left| f_{2,lin}(u_j^n) - f_2(u_j^n) \right| + \frac{1}{2} \cdot \Delta t \cdot \left| fric_{lin}(u_j^n) - fric(u_j^n) \right| .$$

[6] For every error estimator the other variables are "freezed".

[7] In the current implementation the maximal ratio between linearization and discretization error can be prescribed.

Again, for the refinement decision, these error estimates are compared to the corresponding estimated discretization errors.

## 8 Numerical results

For testing purposes, we have implemented our model in the form of a black-box simulator within an approved optimization framework [13]. During the simulation process, the linearized model equations are solved for given boundary conditions and control variables. Gradient information is computed by using difference quotients. After every run of the optimization tool, the error estimators are evaluated and the linearizations are locally refined where necessary.

### 8.1 Example 1 - single compressor

Our first test network consists of two pipes and one compressor as shown in figure 2. The compressor constants are $c_P = 1.10 \cdot 10^2$, $c_F = 3.70 \cdot 10^{-2}$ and $\gamma = 1.4$. Both pipes are 50 km long with a diameter of 1 m and a roughness of 0.01 mm.

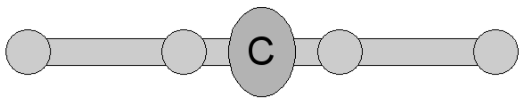


Figure 2: Network with a single compressor.

For the space-time-discretization each pipe is divided into $Nx$ parts and the simulation time of four hours into $Nt$ parts. As initial conditions we use a constant flux of $1.5 \cdot 10^6 \frac{m^3}{h}$ and the pressure linearly decreases from 65bar to 58.2bar in the first pipe and from 58.2bar to 50.5bar in the second. The pressure is kept constant at the beginning of the first pipe and the flux is kept constant at the end of the second pipe.

The target pressure values at the sink (end of the second pipe) are shown in figure 3. The compressor power is set to zero at the beginning and can be configured in every second time step. Intermediate values are computed via linear interpolation.

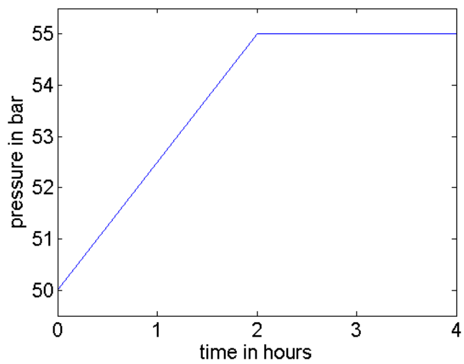


Figure 3: Target pressure at the sink.

For the twodimensional linearizations (functions $f_2$ and *fric*), we start with a coarse regular triangulation as shown in figure 1 ($p_{min} = 49$bar, $p_{max} = 69$bar, $q_{min} = 0\frac{m^3}{h}$, $q_{max} = 2.5 \cdot 10^6 \frac{m^3}{h}$). For the linearizations of the compressor equations, we use the following prism as feasible domain which is initially divided into six tetrahedrons:

$$\begin{aligned}(p_{in}, p_{out}, q_{in}) \quad = \quad & (54\text{bar}, 54\text{bar}, 0\frac{m^3}{h}) \\ & +\lambda_1 \cdot (6\text{bar}, 6\text{bar}, 0\frac{m^3}{h}) \\ & +\lambda_2 \cdot (0\text{bar}, 10\text{bar}, 0\frac{m^3}{h}) \\ & +\lambda_3 \cdot (0\text{bar}, 0\text{bar}, 2.5 \cdot 10^6 \frac{m^3}{h})\end{aligned}$$

with $\lambda_i \in [0, 1]$. Concerning the refinement strategy, a maximal linearization error of 25% of the discretization error is allowed.

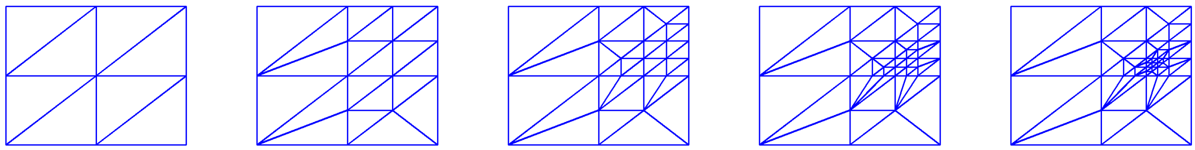

Figure 4: Course of a 2D-triangulation.

| | Nx | Nt | fuel gas | target violation |
|---|---|---|---|---|
| linearized, step 1 | 2 | 4 | 7062 | 0.018 |
| linearized, step 2 | 2 | 4 | 4651 | 0.319 |
| linearized, step 3 | 2 | 4 | 4351 | 0.478 |
| linearized, step 4 | 2 | 4 | 4362 | 0.401 |
| linearized, step 5 | 2 | 4 | 4336 | 0.416 |
| linearized, step 6 | 2 | 4 | 4335 | 0.413 |
| linearized, step 7 | 2 | 4 | 4334 | 0.414 |
| linearized, step 8 | 2 | 4 | 4334 | 0.414 |
| nonlinear | 2 | 4 | 4334 | 0.414 |
| nonlinear | 4 | 8 | 3998 | 0.279 |
| nonlinear | 8 | 16 | 3941 | 0.137 |
| nonlinear | 16 | 32 | 3918 | 0.070 |

Table 1: Optimization results for the first example.

The results of the optimization process are listed in table 1. The fuel gas consumption and the target violation are computed using the "optimal" compressor configuration and a very fine discretization (Nx = 128, Nt = 256). Already after a few refinement steps the solution of the linearized model is very close to the solution of the nonlinear model with the same space-time-discretization. After seven refinement steps there is no further refinement of any triangulation. Figure 4 shows the refinement steps for one of the 2D-triangulations.

## 8.2 Example 2 - two compressors

Our second test network consists of five pipes and two compressors as shown in figure 5. The compressor constants are the same as in the first example. All pipes are 50 km long with a diameter of 1 m and a roughness of 0.005 mm.

The initial conditions are listed in table 2. For the whole simulation time of eight hours, the pressure is kept constant at the beginning of the first pipe and the flux is kept constant at both sinks $S_1$ and $S_2$.

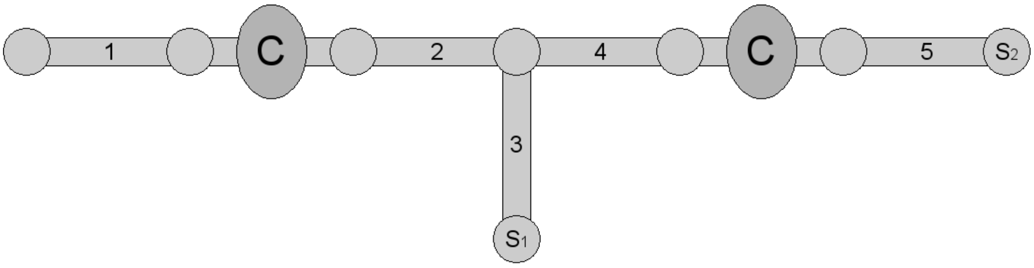


Figure 5: Network with two compressors.

| | pipe 1 | pipe 2 | pipe 3 | pipe 4 | pipe 5 |
|---|---|---|---|---|---|
| $p(0,0)$ in bar | 65.0 | 58.7 | 51.7 | 51.7 | 48.2 |
| $p(L,0)$ in bar | 58.7 | 51.7 | 50.8 | 48.2 | 44.4 |
| $q(x,0)$ in $10^6 \frac{m^3}{h}$ | 1.5 | 1.5 | 0.5 | 1.0 | 1.0 |

Table 2: Initial conditions for the second example.

The target pressure values at the sinks are shown in figure 6. The feasible domain for the initial triangulations has been adapted. Table 3 shows the results of the optimization process. Similar to the first example, the solution of the linearized model is close to the solution of the nonlinear model after a few refinement steps.

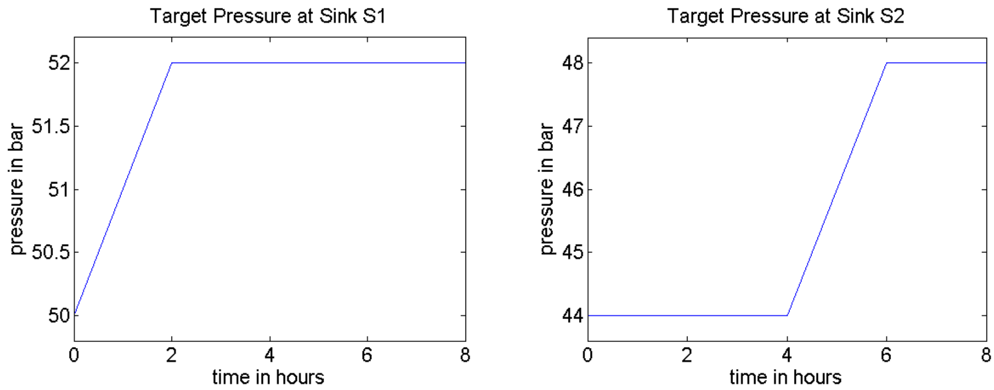


Figure 6: Target pressure at the sinks.

| | Nx | Nt | fuel gas | target violation |
|---|---|---|---|---|
| linearized, step 1 | 2 | 8 | 17569 | 0.000 |
| linearized, step 2 | 2 | 8 | 10085 | 0.000 |
| linearized, step 3 | 2 | 8 | 6282 | 0.000 |
| linearized, step 4 | 2 | 8 | 5314 | 0.123 |
| linearized, step 5 | 2 | 8 | 5222 | 0.163 |
| linearized, step 6 | 2 | 8 | 5205 | 0.179 |
| linearized, step $\geq 7$ | 2 | 8 | 5204 | 0.180 |
| nonlinear | 2 | 8 | 5192 | 0.195 |
| nonlinear | 4 | 16 | 4912 | 0.194 |
| nonlinear | 8 | 32 | 4815 | 0.130 |

Table 3: Optimization results for the second example.

## 9 Conclusion

We have presented an adaptive approach for the linearization process in solving optimal control problems for gas networks using methods provided by discrete optimization. For testing purposes, we have implemented our model in the form of a black-box simulator within an approved optimization framework and we have examined two optimization scenarios. Already after a few refinement steps, the solution of the linearized model is very close to the solution of the nonlinear model with the same space-time-discretization in both test examples. Furthermore, the number of refinements in every step decreased as the solution of the linearized model approached the nonlinear one. Altogether, the first results of our approach are promising and we intend to apply the same techniques for optimal control problems in water distribution networks.

As a next step, we aim to implement our adaptive linearization algorithm in a mixed integer programming framework. Moreover, we intend to investigate how far refinement in space and time is necessary for scenarios with practical relevance.

## Acknowledgements

This paper results from a cooperative scientific project with A. Martin (TU Darmstadt) and G. Leugering (U Erlangen-Nürnberg), supported by the DFG (German Research Foundation).

## Bibliography


[1] M. K. Banda, M. Herty and A. Klar. *Gas flow in pipeline networks*. NHM, Vol. 1, 2006.

[2] M. K. Banda, M. Herty and A. Klar. *Coupling conditions for gas networks governed by the isothermal Euler equations*. NHM, Vol. 1, 2006.

[3] A. Chalabi. *On convergence of numerical schemes for hyperbolic conservation laws with stiff source terms*. Mathematics of computation, Vol. 66, number 218, pp. 527-545, 1997.

[4] Crandall and Madja. *Monotone difference approximations for scalar conservation laws*. Mathematics of computation, Vol. 34, pp. 1-21, 1980.

[5] T. A. Davis. *Direct Methods for Sparse Linear Systems*. SIAM, Philadelphia, 2006.

[6] M. Herty. *Modeling, Simulation and Optimization of Gas Networks with Compressors*. Accepted to NHM, 2006.

[7] S. Karni, A. Kurganov and G. Petrova. *A smoothness indicator for adaptive algorithms for hyperbolic systems.* Journal of Computational Physics 178, pp. 323-341, 2002.

[8] S. Karni and A. Kurganov. *Local error analysis for approximate solutions of hyperbolic conservation laws.* Advances in Computational Mathematics 22, pp. 79-99, 2005.

[9] S. N. Kružkov. *First order quasi-linear equations in several independent variables.* Math. USSR-Sb., 10, (1970), pp. 217-243. MR 42:2159.

[10] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems.* Cambridge University press, pp. 244-252, 2002.

[11] A. Martin, M. Möller and S. Moritz. *Mixed integer models for the stationary case of gas network optimization.* Mathematical Programming, 105:563-582, 2006.

[12] S. Moritz. *A Mixed Integer Approach for the Transient Case of Gas Network Optimization.* PhD thesis, Technical University of Darmstadt, 2006.

[13] P. Spellucci. *donlp2: do nonlinear programming.* Optimization code.