Directed Studies: Notes

S. Gerberding

January 10, 2026

Contents

Chapter 1

From *Finite Elements III*

1.1 Preliminaries

Before we jump into the theory, it is essential that we understand the notation. The symbols we use, especially in higher dimensions, is actually hiding a lot of machinery, and thus it becomes imperatives that we understand what each symbol represents. Let us first set the scene:

$$\begin{aligned} &D \subset \mathbb{R}^d \\ T := [0, \hat{T}) \subset \mathbb{R}_+ &\text{ (or all of } \mathbb{R}_+ \\ &u : D \times T \to \mathbb{R}^m \\ &f : \mathbb{R}^m \to \mathbb{R}^{m \times d} \text{ So the range of } f \text{ is matrices of the size "size of range of } u\text{" by "space domain"} \end{aligned}$$

To jump the shark a bit, the equation we care about is:

$$\partial_t u + \nabla \cdot f(u) = 0 \tag{1.1}$$

For example, what do we mean by the divergence of a matrix? Let us now break down the various cases, in particular, the difference

Case1: $D \subset \mathbb{R}$; $u(x,t) \in \mathbb{R}$ (Purely Scalar Case)
In this case, $\partial_t u$ is scalar valued, $f : \mathbb{R} \to \mathbb{R} = \mathbb{R}^{1\times 1}$. Thus, $\nabla \cdot f(u) = \mathrm{d}x f(u) \in \mathbb{R}$.

Case 2: $D \subset \mathbb{R}^d$; $u(x,t) \in \mathbb{R}$.
Thus, $\partial_t u \in \mathbb{R}$, but $f(u) \in \mathbb{R}^d$, and so $\nabla f(u) = \frac{\partial f(u)}{\partial x_1} + ... + \frac{\partial f(u)}{\partial x_d}$, or the classical divergence.

Case 3: $D \subset \mathbb{R}^d$; $u(x,t) \in \mathbb{R}^m$.
To be clear: $u(x,t) \in \mathbb{R}^m$, $f(u) \in \mathbb{R}^{m\times d}$. (So what is the divergence of a matrix mean?) Well (by definition):

$$\nabla \cdot f(u) = \sum_{j=1}^{d} \partial x_j \begin{bmatrix} f_{1j}(u) \\ f_{2j}(u) \\ \vdots \\ f_{mj}(u) \end{bmatrix} = \sum_{j=1}^{d} \begin{bmatrix} \partial x_j f_{1j}(u) \\ \partial x_j f_{2j}(u) \\ \vdots \\ \partial x_j f_{mj}(u) \end{bmatrix} = \sum_{j=1}^{d} \begin{bmatrix} \frac{\partial f_{1j}}{\partial u_1}\frac{\partial u_1}{\partial x_j} + ... + \frac{\partial f_{1j}}{\partial u_m}\frac{\partial u_m}{\partial x_j} \\ \frac{\partial f_{2j}}{\partial u_1}\frac{\partial u_1}{\partial x_j} + ... + \frac{\partial f_{2j}}{\partial u_m}\frac{\partial u_m}{\partial x_j} \\ \vdots \\ \frac{\partial f_{1m}}{\partial u_1}\frac{\partial u_1}{\partial x_j} + ... + \frac{\partial f_{1m}}{\partial u_m}\frac{\partial u_m}{\partial x_j} \end{bmatrix} = \sum_{j=1}^{d} \begin{bmatrix} \sum_{k=1}^{m} \frac{\partial f_{1j}}{\partial u_k}\frac{\partial u_k}{\partial x_j} \\ \vdots \\ \sum_{k=1}^{m} \frac{f_{mj}}{u_k}\frac{\partial u_k}{\partial x_j} \end{bmatrix}$$

Therefore, $\nabla \cdot f(u)_p = \sum_{j=1}^{d}\sum_{k=1}^{m} \frac{\partial f_p j}{\partial u_k}\frac{\partial u_k}{\partial x_j}$

Next we prove a theorem that becomes the hidden work horse later on for weak solutions, the Fundamental Theorem (or Lemma) of Variational Calculus. We present and prove in full generality.

Theorem 1.1 (Fundamental Theorem of Variational Calculus)**.** *Let $f \in L^1_{loc}(\mathbb{R}^d)$. If $\int f\phi \, \mathrm{d}x = 0$ for all $\phi \in C_0^\infty(\mathbb{R}^d)$, then $f = 0$ a.e.*

Proof. Without loss of generality, suppose $f \geq 0$ (other wise split in positive and negative parts). Let K be compact such that $m(K) > 0$. Choose $\phi \in C_0^\infty(\mathbb{R}^d)$ such that $K \subset supp(\phi)$. (Choose a bump function if necessary). Then, $\int_K f\phi \leq \int_{supp(\phi)} f\phi = \int f\phi = 0$. Thus, (since $f\phi$ is positive) $f\phi = 0$ a.e on K, but since $\phi \neq 0$ on K, $f = 0$ a.e. on K. Thus, for every compact K, $f|_K = 0$ a.e. Next, take the collection of closed balls, $\overline{B(n,)} =: B_n$. These are compact, so $f|_{B_n} = 0$ a.e. for all $n \in \mathbb{N}$. Now, $\mathbb{R}^d = \cup_1^\infty B_n$. Thus, (since f is positive) $f\chi_{B_n} \uparrow f$. Thus, by the Monotone Convergence Theorem, $0 = \int_{B_n} f \to \int f = 0$. Thus, $\int f = 0$, and since f is positive, we deduce that $f = 0$ a.e. □

1.2 Chapter 79: Scalar Conservation Equations

Overview:

1. Scalar- conservation equations
2. Weak and entropy solutions

1.2.1 Weak and entropy solution

Model Problem

- $D \subset \mathbb{R}^d$, polyhedron
- $f \in Lip(\mathbb{R}, R^d)$, Lipschitz vector-valued function; the *flux*
- $u_0 \in L^\infty(D)$; initial data
- PDE: The conservation equation

$$\partial_t u + \nabla \cdot f(u) = 0; \quad u(x,0) = u_0(x), \quad (x,t) \in D \times R_+ \tag{1.2}$$

Motivation for terminology:
Let $O \subset D$ be open. Suppose u solves 1.2. Then:

$$\begin{aligned}
\partial_t u + \nabla \cdot f(u) = 0 \quad &\implies \\
\int_O \partial_t u + \nabla \cdot f(u) \mathrm{d}x &= 0 \\
\partial_t \int_O u \mathrm{d}x + \int_O \nabla \cdot f(u) \mathrm{d}x &= 0 \\
\partial_t \int_O u \mathrm{d}x + \int_{\partial O} f(u) \cdot n ds &= 0 \text{ by Divergence Theorem} \\
\partial_t \int_O u \mathrm{d}x &= -\int_{\partial O} f(u) \cdot n ds.
\end{aligned}$$

That is, the only change in mass inside the region is what leaves the region through the boundary, ie., *no* spontaneous creation or destruction of mass. Hence conservation law. In particular, if $\int_{\partial O} f(u) \cdot n ds = 0$, then the mass within the boundary remains constant.

Here are some important examples:

Example 1.2 (Linear Transport)**.** PDE: $\partial_t u + \beta \nabla u = 0$, where β is some constant vector field, ie, $f(u) = \beta u$. The solution is $u(x,t) = u_0(x - t\beta)$, that is, the data, but shifted.

Example 1.3 (Burger's Equations)**.** Flux $= f(u) = \frac{1}{2}u^2$; PDE $= \partial_t u + \partial_x(\frac{1}{2}u^2) = 0$.

Local existence and loss of smoothness

Definition 1.4. A solution to 1.2 is a *strong solution* over the time interval $[0,T)$ if $u \in C^1(D \times [0,T))$ and u solves 1.2 for all $(x,t) \in D \times [0,T)$.

Remark 1.5. In this setting, we may use the chain rule on 1.2 and recast the problem as:

$$\partial_t u + f'(u) \cdot \nabla u = 0$$

which in a nonlinear transport equation with velocity $f'(u)$.

Method of characteristics

This method comes from more elementary PDE theory, but we go into some better detail. Consider the PDE problem:

$$\begin{cases} \partial_t u + \partial_x f(u) = 0 \\ u(x,0) = u_0(x) \end{cases}.$$

Let us further suppose that $f \in C^1$. Then the PDE becomes:

$$\partial_t u + f'(u)\partial_x u = 0.$$

We define the *characteristic* to be the solution to the following ODE:

$$\begin{cases} \frac{d}{dt}\chi(s,t) = f'(u(\chi(s,t),t)) \\ \chi(s,0) = s \end{cases}.$$

To get a better grasp of this χ, it is related to a path in space time $(x(t),t)$, where $\chi(s,t) = x$. (Here, s is fixed because it is fixed in the ODE problem.) That is, I give you a t, and χ gives me the x.

First: We define the auxiliary function: $\phi(t,s) := u(\chi(s,t),t)$. Then:

$$\partial_t \phi(t,s) = \partial_t u(\chi(s,t),t) = \frac{d}{d\chi} u \frac{d\chi}{dt} + \partial_t u = \partial_x u f'(u) + \partial_t u = 0.$$

Thus, $\phi(t,s)$ is constant with respect to t. Then $\phi(t,s) = \phi(0,s) = u(\chi(s,t),0) = u(s,0) = u_0(s)$. Therefore, $u(\chi(s,t),t) = u_0(s)$. In particular, *the solution is constant along characteristics.* We call this the *implicit solution.*

Second: Using this fact, we deduce:

$$\frac{d\chi(s,t)}{dt} = f'(u(\chi(s,t),t)) = f'(u_0(s)).$$

Thus, our ODE is independent of t, and so solving it we get $\chi(s,t) = f'(u_0(s))t + s$. In particular, the characteristic is simply a line in space time with slope (velocity) $f'(u_0(s))$. We can thus view $\chi(s,t) = x$ to get the full solution.

Example 1.6. Consider the Burgers Equation, with initial data $u_0(x)$ being the hat function:

$$u_0(x) = \begin{cases} 0 & x < -1 \\ 1+x & -1 \le x \le 0 \\ 1-x & 0 < x < 1 \\ 0 & x \ge 1 \end{cases}$$

Now, (as before), we can rewrite the PDE as:

$$u_t + f'(u)u_x = 0.$$

Observe that $f'(u) = u$. We now consider a path though space time $(x(t),t) \subset \mathbb{R}^2$ where $x(0) = x_0$. Now define an auxiliary function $\phi(t) := u(x(t),t)$. Thus:

$$\partial_t \phi = \frac{du}{\mathrm{d}x}\frac{\mathrm{d}x}{dt} + \frac{du}{\mathrm{d}x}. \tag{1.3}$$

Observe how similar this looks to the PDE. Indeed, if we consider the solution to the following ODE:

$$\left\{ \frac{\mathrm{d}x}{dt} = f'(u(x(t),t))u(x(0),t) = u_0(x_0) \right. \quad .$$

Then, 1.3 becomes:

$$\partial_t \phi = \frac{du}{\mathrm{d}x}\frac{\mathrm{d}x}{dt} + \frac{du}{\mathrm{d}x} = f'(u(x(t),t))u_x + u_t = 0. \tag{1.4}$$

Therefore, $\phi(t)$ is constant (along the path). Thus:

$$\phi(t) = \phi(0) = u(x(0),0) = u_0(x_0).$$

We thus conclude that $u(x(t),t) = \phi(t) = u_0(x_0)$. Now we only need to find x in terms of x_0 based on $u_0(x_0)$.

When $x_0 < 1$, $u_0(x_0) = 0$, and thus the point remains where it is. The same is true for $x_0 > 1$. When $-1 \leq x_0 \leq 0$, we have $u_0(x) = x_0 + 1 = \frac{\mathrm{d}x}{dt}$. Thus, solving this ODE for x gives $x = (1+x_0)t + x_0$, so $x_0 = \frac{x-t}{1+t}$. Thus:

$$u(x(t),t) = u_0(x_0) = 1 + x_0 = 1 + \frac{x-1}{t+1} = \frac{1+x}{1+t}.$$

A similar approach gives the result for when $0 < x_0 < 1$.

Example 1.7. Consider another Burgers Equation set up, but this time $u_0(x) = \sin(x)$. As before, $f'(u) = 0$, and we consider a path $(x(t),t)$ with $x(0) = x_0$. Then, we consider an ODE $\frac{\mathrm{d}x}{dt} = f'(u(x(t),t))$. As before, $\frac{\mathrm{d}x}{dt} = u_0(x) = \sin(x_0)$. Thus, solving this ODE gives $x = \sin(x_0)t + x_0$. This gives an implicit solution for x_0. In this case, we can use Newton's Method (or Euler's Method (?)) to find a solution. Finally, as in the first exmaple, we have $u(x(t),t) = u_0(x_0) = \sin(x_0) =$ numerical result.

Proposition 1.8. Assume that f is class C^2 and that u_0 is class C^1 and $\inf_{s\in\mathbb{R}} \min(f''(u_0(x))u_0'(s),)) > -\infty$. Then 1.2 has a unique strong solution over the time interval.

Weak Solutions

It may very well be the case (if not frequently so) that the initial data is not smooth, ie, there is a jump discontinuity or some edge, or smoothness is lost in some finite amount of time (that is, the solution starts smooth, but then breaks in finite times). We would still like to have some notion of solution, but to do so, we must drop our insistence that u_0 be in C^1 (among other possible assumptions). We therefore introduce the notion of *weak solution* (which is related to, in spirit, the weak derivative, insofar as the weak derivative is an extremely simple kind of PDE). To develop the weak solution, we integrate the PDE (in both time and space) against some smooth (C^1), compactly supported test function and use integration by parts. The general starting form is:

$$\int_D \partial_t u \cdot \phi + \nabla \cdot f(u)\phi \ \mathrm{d}x = 0 \tag{1.5}$$

$$\int_D \partial_t u \phi \mathrm{d}x + \int_D \nabla \cdot f(u) \cdot \phi \ \mathrm{d}x = 0 \tag{1.6}$$

$$\tag{1.7}$$

We make the computation slowly, integrating in space first. Let us first suppose that we are in the purely scalar case. Let $\phi \in C_0^1(D \times \mathbb{R}_+; \mathbb{R})$.

$$\int_D \partial_t u\phi + \partial_x f(u)\phi \, \mathrm{d}x = 0 \tag{1.8}$$

$$\int_D \partial_t u\phi \mathrm{d}x + \int_D \partial_x f(u)\phi \, \mathrm{d}x = 0 \tag{1.9}$$

$$\tag{1.10}$$

However, because ϕ is zero on the boundary, by Stoke's Theorem:

$$\int_D \partial_x f(u)\phi \mathrm{d}x = \int_{\partial D} f(u)\phi dx - \int_D f(u)\cdot\partial_x\phi \mathrm{d}x = 0 - \int_D f(u)\cdot\partial_x\phi \mathrm{d}x = -\int_D f(u)\cdot\partial_x\phi \mathrm{d}x.$$

Thus, we can improve 1.6 to:

$$\int_D \partial_t u\phi \mathrm{d}x + \int_D \partial_x f(u)\phi \mathrm{d}x = \int_D \partial_t u\phi \mathrm{d}x - \int_D f(u)\cdot\nabla\phi \mathrm{d}x = 0.$$

Let's now integrate in time, and focus this time on the first term:

$$\begin{aligned}
\int_0^\infty \int_D \partial_t u\phi \mathrm{d}x dt &= \int_D \int_0^\infty \partial_t u\phi dt \mathrm{d}x \text{ by Fubini} \\
&= \int_D \left[u\phi|_0^\infty - \int_0^\infty u\partial_t\phi dt \right] \mathrm{d}x \\
&= \int_D \left[-u(0,x)\phi(0,x) - \int_0^\infty u\partial_t\phi dt \right] \mathrm{d}x \text{ since } \phi \text{ is compactly supported} \\
&= -\int_D u_0(x)\phi(0,x)\mathrm{d}x - \int_D \int_0^\infty u\partial_t\phi dt \mathrm{d}x.
\end{aligned}$$

Thus, putting both calculations together, we conclude:

$$\int_0^\infty \int_D \partial_t u\phi + \partial_x f(u)\phi \mathrm{d}x dt = -\int_D u_0(x)\phi(0,x)\mathrm{d}x - \int_D \int_0^\infty u\partial_t\phi dt \mathrm{d}x - \int_0^\infty \int_D f(u)\partial_x\phi \mathrm{d}x dt = 0.$$

Thus, taking negative gives the result for this case.

When $u(x,t) \in \mathbb{R}^m$, an easy extension is as follows: let $\phi \in C_0^1(D \times \mathbb{R}_+; \mathbb{R}^m)$. Then a similar calculation (using more general theorems) gives:

$$\int_D u_0(x)\cdot\phi(0,x)\mathrm{d}x + \int_D \int_0^\infty \left[u\cdot\partial_t\phi + f(u)\cdot\partial_x\phi \right] \, dt\mathrm{d}x = 0 \tag{1.11}$$

where $\cdot$ is the $\mathbb{R}^m$ dot product.

In the full general case ($u(x,t) \in \mathbb{R}^m, \ D \subset \mathbb{R}^d$): let $\phi_0^1(D \times \mathbb{R}_+, \mathbb{R}^m)$:

$$\int_D u_0(x)\cdot\phi(0,x) \, \mathrm{d}x + \int_D \int_0^\infty \left[u\cdot\partial_t\phi + f(u) : \nabla\phi \right] \, \mathrm{d}x dt = 0 \tag{1.12}$$

where $\cdot$ (again) is the $\mathbb{R}^m$ dot product, but : is the double concatination.

Remark 1.9 (A more general weak solution)**.** Suppose that our time domain was restricted to some $[0,T]$. Then, the weak solution is u such that the following holds for all $\phi \in C_0^1[D \times [0,T))$ (the computation is similar to the above):

$$\int_0^T \int_D u\partial_t\phi + f(u)\cdot\nabla\phi \, \mathrm{d}x dt + \int_D u_0(x)\phi(x,0) - u(x,T)\phi(x,T) \, \mathrm{d}x = 0.$$

Definition 1.10. A function $u \in L^\infty_{loc}(D \times \mathbb{R}_+)$ is a *weak solution* if it satisfies 1.12 for all $\phi \in C^1_0(D \times \mathbb{R}_+; \mathbb{R}^m)$.

Example 1.11. Let $D = \mathbb{R}^d$, $u_0 \in L^\infty_{loc}(D)$. Let us show that $u(x,t) := u_0(x - \beta t)$ is indeed a weak solution to the linear transport equations. Recall that in the linear transport equation, $f(u) = \beta u$ for some constant β. We show that the left hand side of 1.12 does equal 0. Consider:

$$I := \int_0^\infty \int_D u\partial_t\phi(x,t) + (u\beta) \cdot \nabla\phi(x,t)\mathrm{d}x dt$$

Since $u(x,t) = u_0(x - \beta t)$ (by assumption), we achieve:

$$I = \int_0^\infty \int_D u_0(x - \beta t)\partial_t\phi(x,t) + (u_0(x - \beta t)\beta) \cdot \nabla\phi(x,t)\mathrm{d}x dt.$$

Now make a change of variable: $x' = x - \beta t$, so $\mathrm{d}x' = \mathrm{d}x$. Then:

$$\int_0^\infty \int_D u_0(x')\phi(x' + \beta t, t) + (u_0(x')\beta) \cdot \nabla\phi(x' + \beta t)\mathrm{d}x' dt.$$

We would like to make things simpler, so let $\psi(x', t) := \phi(x' + \beta t, t)$. Then:

$$\partial_t\psi(x', t) = \nabla\phi(x' + \beta t, t)\beta + \partial_t\phi(x' + \beta t, t).$$

We therefore deduce:

$$\begin{aligned}
\int_0^\infty \int_D u_0(x')\partial_t\phi(x' + \beta t, t) + (u_0(x')\beta) \cdot \nabla\phi(x' + \beta t)\mathrm{d}x' dt &= \int_0^\infty \int_D u_0(x')\partial_t\psi(x', t)\mathrm{d}x' dt \\
&= \int_D u_0(x') \int_0^\infty \partial_t\psi(x', t) dt \mathrm{d}x' \\
&= \int_D u_0(x') \left[-\psi(x', 0)\right] \mathrm{d}x \text{ since } \psi \text{ is compactly supported} \\
&= \int_D -u_0(x')\psi(x', 0) = \int_D -u_0(x')\phi(x', 0)\mathrm{d}x'.
\end{aligned}$$

Bringing the right hand side over shows that $u(x,t) := u_0(x - \beta t)$ is indeed a weak solution.

Remark 1.12. In general, there are *infinitely many* weak solutions to a given problem. To achieve uniqueness, we must find a way to separate a physically realistic weak solution from nonsense solutions.

Existence and uniqueness

Definition 1.13. A function u is a *physically relevant solution* to 1.2 if it is a weak solution and if it is the limit (in some appropriate topology) of the unique solution to the following perturbed problem as $\epsilon \to 0$:

$$\partial_t u_\epsilon + \nabla \cdot f(u_\epsilon) - \epsilon\Delta u_\epsilon = 0; \quad u_\epsilon(x, 0) = u_0(x); \quad (x,t) \in D \times \mathbb{R}_+. \tag{1.13}$$

We call u_ϵ the *viscous regularization of* u.

Remark 1.14. One way to look at this approximation is that we've allowed the solution more wiggle room, and (I believe) this problem is much much simpler to solve and guarantee a unique solution exists).

Remark 1.15. It was shown that u be a limit of the above solutions, ie, $\lim_{\epsilon\to 0} ||u - u_\epsilon||_{L^1} = 0$ is equivalent to u satisfying entropy inequalities:

$$\partial_t\eta(u) + \nabla \cdot q(u) \leq 0$$

Theorem 1.16. *Let $f \in Lip(\mathbb{R}, \mathbb{R}^d)$ and $u_0 \in L^\infty(D)$. There is a unique entropy solution to 1.2, is, there is a u that is a weak solution and satisfies:*

$$-\int_0^\infty \int_D \eta(n)\partial_t\phi + q(u)\cdot\nabla\phi \mathrm{d}x dt - \int_D \phi(x,0)\eta(u_0)\mathrm{d}x \leq 0$$

for all entropy pairs (η, q) and all compactly support $\phi \in C_0^1(D \times \mathbb{R}_+, \mathbb{R}_+)$.

Theorem 1.17. *Let $u_{min} = ess\inf_{x\in D} u_0(x)$ and $u_{max} = ess\sup_{x\in D} u_0(x)$. Then the entropy solution satisfies the maximum principle:*

$$u(x,t) \in [u_{min}, u_{max}].$$

1.2.2 Riemann Problem

The Riemann Problem (first proposed, obviously, by Riemann) is a classic conservation law problem based upon a rather simple initial condition. One prerequisite is that the flux is Lipschitz.

1D Riemann Problem

We consider the 1D Riemann problem:

$$\begin{cases} \partial_t u + \partial_x f(u) = 0 \\ u(x,0) = \begin{cases} u_L & x < 0 \\ u_R & x > 0 \end{cases} \end{cases}. \tag{1.14}$$

Case 1: $u_L < u_R$. We first suppose that $f \in C^2$ and f is convex, ie, $f'' > 0$ on the domain $[u_L, u_r]$. That is, f' is monotonically increasing. We use the method of characteristics.

First suppose $x_0 < 0$, so $u_0(x_s) = u_L$. Then the charactersistic becomes:

$$\chi(t, x_0) = x_0 + tf'(u_L).$$

Observe that since u_L is constant, the characteristics are all parallel (for $x_0 < 0$). Thus, the implicit solution is:

$$u(\chi(t,x_0),t) = u_0(\chi(t,x_0) - f'(u_L)t) = u_0(x_0) = u_L.$$

Thus, we conclude this is only valid when $\chi(t,x_0) - f'(u_L)t < 0$, or $\chi(t,x_0)/t < f'(u_L)$. We can thus recast this as $u(x,t) = u_L$ when $x/t < f'(u_L)$.

Likewise, if we suppose $x_0 > 0$, then $u_0(x_0) = u_R$, so $\chi(t,x_0) = x_0 + tf'(u_R)$. Thus:

$$u(\chi(t,x_0),t) = u_0(x_0) = u_0(\chi(t,x_0) - tf'(u_R)) = u_R.$$

A similar argument as above gives then that $u(x,t) = u_R$ when $x/t > f'(u_R)$.

Since f' was monotone increasing and $u_L < u_R$, everything is perfectly well-defined thus far. The question now becomes what about $f'(u_L) < x/t < f'(u_R)$? The solution appears to depend upon x/t, ie, $u(x,t) = w(x/t)$ where $w : \mathbb{R} \to \mathbb{R}$. If we substitute w into (1.14), we have:

$$\partial_t w + f'(w)\partial_x w = -\frac{x}{t^2}w'(x/t) + w'(x/t)\frac{1}{t}f'(w).$$

If $x/t = f'(w(x/t))$, then:

$$-\frac{x}{t^2}w'(x/t) + \frac{1}{t}w'(x/t)f'(w) = 0.$$

Thus, $u(x,t) = w(x/t)$ solves the problem iff $x/t = f'(w(x/t))$. (Note that the above argument holds for any Riemann problem regardless of the initial conditions or f). But we know that f' is continuous

and monotone increasing, and so $(f')^{-1}$ exists on $[f'(u_L), f'(u_R)]$. Therefore, $w(x/t) = (f')^{-1}(x/t)$ for $f'(u_L) \leq x/t \leq f'(u_R)$. Thus, the solution is:

$$u(x,t) = \begin{cases} u_L & x/t \leq f'(u_L) \\ (f')^{-1}(x/t) & f'(u_L) \leq x/t \leq f'(u_R) \\ u_R & f'(u_R) \leq x/t \end{cases}. \tag{1.15}$$

But what happens if $f'(u_R) < f'(u_L)$? Then we have a *shock*. In particular, if one looks at the characteristics, then they intersect, and when they intersect, there is a shock. Contrasted with the first case, $f'(u_L) < f'(u_R)$, and so the characteristics never intersect, and thus, we have the *expansion wave* as described above.

Case 2: $u_R < u_L$. This case, we first assume that f is concave rather than convex (still C^2 though). Then a similar argument follows, and the solution is similar.

To summarize: when $u_L < u_R$ and f is convex, the solution is as above. When $u_R < u_L$, f is concave, the solution is as above. When this does not occur, we get a shock wave: $u_L < u_R$, but f is concave or $u_R < u_L$, but f is convex.

Example 1.18 (Burger's Equations)**.** Consider the Riemann problem with Burger's Equations, so $f(u) = \frac{u^2}{2}$. Observe that f is convex everywhere. Since $f'(u) = u$, we have that $(f')^{-1}(u) = u$, and so the solution becomes:

$$u(x,t) = \begin{cases} u_L & x/t < (f')^{-1}(u_L) = u_L \\ x/t & u_L < x/t < u_R \\ u_R & u_R < x/t \end{cases}.$$

Up till now, we have assumed that f is either convex or concave (on a certain interval). What about more generally? We consider the *convex envelope* and *concave envelope*:

$$\bar{f}(v) := \sup\{g(v) : g(z) \leq f(z) \text{ for all } z \text{ and } g \text{ convex}\} \tag{1.16}$$

$$\underline{f}(v) := \inf\{g(v) : g(z) \geq f(z) \text{ for all } z \text{ and } g \text{ concave}\}. \tag{1.17}$$

To get a better grasp of these objects (in particular, the convex envelope, for the concave envelope is similar), suppose we have $u_L < u_R$, but f was concave (say something like $-u^2$). Then, the convex envelope is simply the line connecting $(u_L, f'(u_L))$ and $(u_R, f'(u_R))$. Then (draw the damn picture), the convex envelope is $\bar{f}(u) = \frac{f'(u_L)(u-u_R)}{u_L-u_R} + \frac{f'(u_R)(u-u_L)}{(u_R-u_L)}$. Then, $\bar{f}'(u) = \frac{f(u_L)-f(u_R)}{u_L-u_R}$, which is precisely the shock wave speed from before.

It can be shown that $\bar{f}$ and $\underline{f}$ are convex and concave respectively. Furthermore, it can be shown that, is the case that $u_L < u_R$, the solution is:

$$u(x,t) = \begin{cases} u_L & x/t \leq (\bar{f}')^{-1}(u_L) \\ (\bar{f}')^{-1}(x/t) & \bar{f}'(u_L) < x/t < \bar{f}'(u_R) \\ u_R & \bar{f}'(u_R) < x/t \end{cases}.$$

And in the case $u_R < u_L$, the solution is:

$$u(x,t) = \begin{cases} u_L & x/t \leq (\underline{f}')^{-1}(u_L) \\ (\underline{f}')^{-1}(x/t) & \underline{f}'(u_L) < x/t < \underline{f}'(u_R) \\ u_R & \underline{f}'(u_R) < x/t \end{cases}.$$

We combine this all into a single theorem:

Theorem 1.19 (Riemann Solution)**.** *Assume that the interval $[u_L, u_R]$ can be divided into finitely many subintervals where f has a continuous and bounded second derivative, and where f is strictly convex or strictly concave. The entropy solution to the Riemann Problem is given by:*

$$u(x,t) = \begin{cases} u_L & x/t \le (\bar{f}')^{-1}(u_L) \\ (\bar{f}')^{-1}(x/t) & \bar{f}'(u_L) < x/t < \bar{f}'(u_R) \\ u_R & \bar{f}'(u_R) < x/t \end{cases}. \tag{1.18}$$

if $u_L < u_R$, and $\bar{f}$ must be replaced by $\underline{f}$ if $u_L > u_R$.

Riemann Cone and Averages

Define the following objects:

$$\lambda_L(u_L,u_R) := \begin{cases} \bar{f}'(u_L) & u_L < u_R \\ \underline{f}'(u_L) & u_L > u_R \end{cases} \quad \lambda_R(u_L,u_R) := \begin{cases} \bar{f}'(u_L) & u_L < u_R \\ \underline{f}'(u_R) & u_L > u_R \end{cases}.$$

Observe how the solution depends purely on whether $u_L < u_R$ or $u_R < u_L$ and the relationship x/t shares with them. In particular, the solution is nontrivial only in a certain region. When $u_L < u_R$, then it is nontrivial when $\bar{f}'(u_L) < x/t < \bar{f}'(u_R)$; when $u_R < R_L$, it is nontrivial when $\underline{f}'(u_L) < x/t < \underline{f}'(u_R)$. This define the *Riemann Cone* (see page 333 for a picture.) We can recast this more formally: The area of nontrivial solution occurs in the *Riemann Cone*, or the *Riemann Fan*:

$$C(u_L,u_R) := \{(x,t) \in \mathbb{R} \times \mathbb{R}^+ : \ \lambda_L(u_L,u_R) \le x/t \le \lambda_R(u_L,u_R)\}. \tag{1.19}$$

Definition 1.20. We call the *maximum speed* in the Riemann problem the number $\max|\lambda_R(u_R,u_L)|, |\lambda_L(u_R,u_L)|$. Any number satisfying the inequality

$$\lambda_M(u_L,u_R) \ge \max|\lambda_R(u_R,u_L)|, |\lambda_L(u_R,u_L)|$$

is called an upper bound on the maximum wave speed.

Lemma 1.21. *Let u be an entropy solution to (1.14), (n,q) be an entropy pair, and define the Riemann average as $\bar{u}(t,u_L,u_R) := \int_{-1/2}^{1/2} u(x,t)\ dt$. Let λ_M be any upper bound on the maximum wave speed. The for all $t \in [0, \frac{1}{2\lambda_M}]$,*

$$\bar{u}(t,u_L,u_R) = \frac{1}{2}(u_L+u_R) - t(f(u_R)-f(u_L)), \tag{1.20}$$

$$n(\bar{u}(t,u_L,u_R)) \le \frac{1}{2}(n(u_L)+n(u_R)) - t(q(u_R)-q(u_L)). \tag{1.21}$$

Proof. Consider

$$\partial_t u + \partial_x f(u) = 0$$

and integrate from $-1/2$ to $1/2$ in space and 0 to t in space:

$$\int_{-1/2}^{1/2}\int_0^t \partial_t u + \partial_x f(u)\ d\tau \mathrm{d}x = 0.$$

Direct computation yields:

$$\int_0^t \int_{-1/2}^{1/2} \partial_x f(u) \mathrm{d}x d\tau = \int_0^t f(u(1/2,\tau) - f(u(-1/2,\tau)) d\tau;$$

$$\int_{-1/2}^{1/2}\int_0^t = \int_{1/2}^{1/2} u(t,x)\mathrm{d}x - \int_{-1/2}^{1/2} u(0,x)\mathrm{d}x = \bar{u}(t,u_L,u_R) - \int_{-1/2}^{1/2} u_0(x)\mathrm{d}x.$$

Thus:

$$\bar{(u)}(t,u_L,u_R) - \frac{1}{2}(u_L+u_R) + \int_0^t f(u(1/2,\tau)) - f(u(-1/2,\tau)) d\tau = 0.$$

However by assumption, $\frac{-1}{2\tau} \leq \frac{-1}{2t} \leq -\lambda_M \leq \lambda_L$. Thus (looking at equation 1.18), we have $u(1/2, \tau) = u_R$. Therefore,

$$\bar{u}(t, u_L, u_R) - \frac{1}{2}(u_L + u_R) + \int_0^t f(u_R) - f(u_L) d\tau = 0 \implies \bar{u}(t, u_L, u_R) - \frac{1}{2}(u_L + u_R) + t(f(u_R) - f(u_L)) = 0.$$

This gives 1.37. A similar argument gives:

$$\partial_t n(u) + \partial_x q(u) \leq 0 \implies \int_{-1/2}^{1/2} n(u(x,t))\mathrm{d}x - \frac{1}{2}(n(u_L) + (u_R)) + \frac{1}{2}(q(u_R) - q(u_L)) \leq 0.$$

Jenssen's Inequality, $n(\int_{-1/2}^{1/2} u\mathrm{d}x) \leq \int_{-1/2}^{1/2} n(u(x,t))\mathrm{d}x$ gives the inequality. □

Remark 1.22. The max/min principle implies that u is in the convex hull of (u_L, u_R). Thus, $\bar{(u)} \in$ *convex hull*. Then, for t as in the theorem, the quantity above is in the convex hull.

1.3 Chapter 80: Hyperbolic System

1.3.1 First-order Quasi-linear hyperbolic systems

Set up:

- $m \in \mathbb{N},\ m \neq 0$
- $\mathcal{A} \subset \mathbb{R}^m$, called the *admissible set of states*.
- $D \subset \mathbb{R}^d$
- $\mathbf{A}_l \in Lip(\mathcal{A}; \mathbb{R}^{m\times})$
- $u_0 \in \mathcal{A}$
- PDE:

$$\partial_t u + \sum_{l\in\{1:d\}} \mathbf{A}(u)_l \partial_{x_l} u = 0; \quad u(x,0) = u_0(x). \tag{1.22}$$

Of course, $u(x,t) \in \mathbb{R}^m$. The above system is called *first order quasilinear system*.

Definition 1.23. The system in [1.22] is *hyperbolic* if the matrix $\mathbf{A}(v,n) := \sum_{l\in\{1:d\}} n_l \mathbf{A}_l(v)$ is diagonalizable with real eigenvalues for all $v \in \mathcal{A}$ and any unite vector $n \in \mathbb{R}^d$. (Recall that a matrix is diagonaliable iff the sum of the dimensions of the eigenspaces is the same as the image space.) The system is strictly hyperbolic if, in addition, all eigenvalues are distinct.

Suppose we have a nice domain D and that $\mathbf{A}_l(v) = \mathbf{A}_l$ for all $l \in \{1:d\}$ and $v \in \mathcal{A}$. Further suppose thatthere is some $k = (k_1, ..., k_d)$ and $\mathbf{U}_k \in \mathbb{R}^m$ such that $u_0(x) = \mathbf{U}_k e^{ik\cdot x}$. Let $k := \frac{k}{\|k\|_{l^2}^2}$. Now consider:

$$\mathbf{A}(x) := \frac{1}{||k||_{l_2}}^2 \sum_{l\in\{1:d\}} k_l \mathbf{A}_l.$$

(That is, the sum of the each matrix scaled by k_l.) Then, referring to the above definition, this system is hyperbolic iff $\mathbf{A}$ is diagonalizable with real eigenvalues.

Let us suppose that it is hyperbolic. Then, there exists eigenvalues $\{\lambda_1, ..., \lambda_m\}$ with eigenvalues $\{v_1, ..., v_m\}$ such that $\{v_j\}_{j\in\{1:m\}}$ is a basis. Therefore:

$$U_k = \sum_{j\in\{1:m\}} \alpha_j v_j.$$

By assumption, we conclude that

$$u_0(x) = \sum_{j\in\{1:m\}} \alpha_j v_j e^{ik\cdot x}.$$

Using these calculations, we can actually prove a solution exists:

Lemma 1.24 (Plane wave solution)**.** *Under the above assumptions, the unique solution is:*

$$u(x,t) = \sum_{j\in\{1:m\}} \alpha_j v_j e^{i(k\cdot x - \lambda_j \|k\|_{l^2} t}. \tag{1.23}$$

Proof. We readily observe that the initial conditions are satisfied. Thus we only need to verify the PDE. Linearity of the derivative gives:

$$\partial_t u(x,t) = -i \sum_{j\in\{1:m\}} \alpha_j v_j \lambda_j \|k\|_{l^2} e^{i(k\cdot x - \lambda_j \|k\|_{l^2} t};$$

$$\frac{\partial u(x,t)}{\partial x_l} = i \sum_{j\in\{1:m\}} \alpha_j v_j k_l e^{i(k\cdot x - \lambda_j \|k\|_{l^2} t}.$$

Because (λ_j, v_j) is an eigenpair of $\mathbf{A} := \frac{1}{\|k\|_{l^2}} \sum_{l\in\{1:d\}} \mathbf{A}_l k_l$, we deduce

$$\sum_{l\in\{1:d\}} \mathbf{A}_l k_l v_j = \|k\|_{l^2} v_j \text{ for all } j \in \{1:m\}.$$

Therefore:

$$\sum_{l\in\{1:d\}} \mathbf{A}_l \frac{\partial u}{\partial x_l} = i \sum_{j\in\{1:m\}} \|k\|_{l^2} \lambda_j v_j e^{i(k\cdot x - \lambda_j \|k\|_{l^2} t}.$$

And finally, we conclude that $\partial_t u + \sum_{l\in\{1:d\}} \mathbf{A}_l \frac{\partial u(x,t)}{\partial x_l} = 0$, and the proof is complete. □

Note 1.25. The idea is that the solution to 1.3.2 is a planar wave. Then, as we will see in the Riemann problem, we can pick a direction using a unit vector n and recreate a 1D Riemann problem that tells us about the solution along that line.

1.3.2 Hyperbolic system in conservative form

Consider our typical conservation system:

$$\partial_t u + \nabla \cdot \mathbf{f}(\mathbf{u}) = 0.$$

Recall from the preliminary section that this can be rewritten as:

$$\partial_t u + \sum_{l\in\{1:d\}} \begin{bmatrix} \sum_{j\in\{1:m\}} \frac{\partial f_{1l}}{\partial u_j} \frac{\partial u_j}{\partial x_l} \\ \vdots \\ \sum_{j\in\{1:m\}} \frac{\partial f_{ml}}{\partial u_j} \frac{\partial u_j}{\partial x_l} \end{bmatrix} = 0.$$

For $l \in \{1:d\}$, observe

$$\begin{bmatrix} \sum_{j\in\{1:m\}} \frac{\partial f_{1l}}{\partial u_j} \frac{\partial u_j}{\partial x_l} \\ \vdots \\ \sum_{j\in\{1:m\}} \frac{\partial f_{ml}}{\partial u_j} \frac{\partial u_j}{\partial x_l} \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{\partial f_{1l}}{\partial u_1} & \cdots & \frac{\partial f_{1l}}{\partial u_m} \\ \vdots & & \vdots \\ \frac{\partial f_{ml}}{\partial u_1} & \cdots & \frac{\partial f_{ml}}{\partial u_m} \end{bmatrix}}_{\mathbf{A}_l(u)} \underbrace{\begin{bmatrix} \frac{\partial u_1}{\partial x_l} \\ \vdots \\ \frac{\partial u_m}{\partial x_l} \end{bmatrix}}_{\partial x_l u}.$$

Therefore, the conservation equation can be written as:

$$\partial_t u + \sum_{l\in\{1:d\}} \mathbf{A}_l(u)\partial_{x_l} u. \tag{1.24}$$

Following the definition above, we conclude that the conservation system is hyperbolic iff the matrix

$$(\mathbf{A}(v,n))_{ij} := \sum_{l\in\{1:d\}} n_l \partial_{v_j} f_{il}(v), \quad \text{for } i,j \in \{1:m\}$$

is diagonalizable over $\mathbb{R}$.

Definition 1.26 (Weak Solution)**.** We say that $u \in L^\infty_{loc}(D\times\mathbb{R}^+,\mathbb{R}^m)$ is a weak solution to [1.3.2] if for all $\phi \in C^1_0(D\times\mathbb{R}^+,\mathbb{R}^m)$, we have

$$\int_0^\infty \int_D (u\cdot\partial_t\phi + f(u) : \nabla\phi)\ \mathrm{d}x dt + \int_D \phi(x,0)\cdot u_0(x)\ \mathrm{d}x = 0. \tag{1.25}$$

(See preliminary calculations for derivation.)

Remark 1.27. As in the scalar case, while the notion of weak solutions opens many doors, there are too many door, and uniqueness becomes a problem.

1.3.3 Entropy for Hyperbolic Systems

As in the scalar case, another way to discuss/ address uniqueness of a solution is to invoke entropy. As before, we have entropy inequalities, and entropy solutions give us physically relevant solutions.

Definition 1.28. We say that (η, q) is an *entropy pair* for a conservative system (as in [1.3.2]) if the functions are such that :

1. $\eta \in C^1(\mathcal{A},\mathbb{R})$ is convex;
2. $q \in C^1(\mathcal{A},\mathbb{R}^d)$ is such that:

$$\partial v_j q_k(v) = \sum_{i\in\{1:m\}} \partial v_i \eta(v)\ \partial v_j f_{i,k}(v) \text{for all } v\in\mathcal{A}.. \tag{1.26}$$

In particular, given an entropy pair (η, q), we can choose a physically relevant solutions that satisfies the following inequality holds for all $\phi \in C^\infty_0(D\times\mathbb{R}^+,\mathbb{R}^+)$:

$$-\int_0^\infty \int_D (\eta(u)\partial_t\phi + q(u)\nabla\cdot\phi)\ \mathrm{d}x dt - \int_D \phi(x,0)\eta(u_0)\ \mathrm{d}x \le 0. \text{for all } \phi\in C^\infty_0(D\times\mathbb{R}^+,\mathbb{R}^+) \tag{1.27}$$

We now show that if u satisfies [**??**], then it satisfies $\partial_t\eta(u) + \nabla\cdot q(u) \le 0$ a.e. (This is good practice) Consider:

$$\begin{aligned}
-\int_D\int_0^\infty \eta(u)\phi_t\ dt\mathrm{d}x &= -\int_D \left[\eta(u)\phi|_0^\infty - \int_0^\infty \eta(u)_t\phi dt\right]\ \mathrm{d}x \\
&= -\int_D -\eta(u_0(x)\phi(0,x) - \int_0^\infty \eta(u)_t\phi\ \mathrm{d}x dt \\
&= \int_D \eta(u_0(x)\phi(0,x)\mathrm{d}x + \int_D\int_0^\infty \eta(u)_t\phi\ dt\mathrm{d}x.
\end{aligned}$$

$$
\begin{aligned}
-\int_D \int_0^\infty q(u)\nabla\cdot\phi \; dt\mathrm{d}x &= \int_0^\infty -\left[\int_{\partial D} \eta\cdot q(u)\phi \; \mathrm{d}x - \int_D \nabla q(u)\phi \; \mathrm{d}x\right] dt \\
&= -\int_0^\infty -\int_D \nabla q(u)\phi \; \mathrm{d}x dt \\
&= \int_0^\infty \int_D \nabla q(u)\phi \; \mathrm{d}x dt.
\end{aligned}
$$

Combining these calculations together, we deduce

$$
\int_D \eta(u_0(x))\phi(0,x) \; \mathrm{d}x + \int_D \int_0^\infty \eta(u)_t\phi \; dt\mathrm{d}x + \int_0^\infty \int_D \nabla q(u)\phi \; \mathrm{d}x dt - \int_D \eta(u_0(x))\phi(0,x) \; \mathrm{d}x \leq 0.
$$

Therefore, we have:

$$
\int_D \int_0^\infty \eta(u)_t\phi + \nabla q(u)\phi \; dt\mathrm{d}x = \int_D \int_0^\infty \left(\eta(u)_t + \nabla q(u)\right)\phi \; dt\mathrm{d}x \leq 0.
$$

And thus by the Fundamental Theorem of Variational Calculus,

$$
\eta(u)_t + \nabla q(u) \leq 0. \tag{1.28}
$$

Proposition 1.29. If a weak solution u to [1.24], that is, u satisfies [1.25], is smooth, then the inequalities in [1.28] are in fact equalities.

Proof. Since u is smooth, we may apply the chain rule:

$$
\partial_t \eta(u) = \sum_{j\in\{1:m\}} \frac{\partial \eta}{\partial u_j}\frac{\partial u_j}{\partial t};
$$

$$
\nabla\cdot q(u) = \sum_{i\in\{1:d\}} \frac{\partial q_i(u)}{\partial x_i} = \sum_{i\in\{1:D\}} \sum_{k\in\{1:m\}} \frac{\partial q_i}{\partial u_k}\frac{\partial u_k}{\partial x_i}.
$$

Therefore:

$$
\begin{aligned}
\partial_t\eta(u) + \nabla q(u) &= \sum_{j\in\{1:m\}} \frac{\partial \eta}{\partial u_j}\frac{\partial u_j}{\partial t} + \sum_{i\in\{1:d\}} \sum_{k\in\{1:m\}} \frac{\partial q_i}{\partial u_k}\frac{\partial u_k}{\partial x_i} \\
&= \sum_{j\in\{1:m\}} \frac{\partial \eta}{\partial u_j}\frac{\partial u_j}{\partial t} + \sum_{i\in\{1:d\}} \sum_{k\in\{1:m\}} \sum_{l\in\{1:m\}} \frac{\partial \eta}{\partial u_l}\frac{\partial f_{i,k}}{\partial u_k}\frac{\partial u_k}{\partial x_i} \quad \text{by the definition of entropy pair} \\
&= \sum_{j\in\{1:m\}} \frac{\partial \eta}{\partial u_j}\left(\underbrace{\frac{\partial u_j}{\partial t} + \sum_{i\in\{1:d\}} \sum_{k\in\{1:m\}} \frac{\partial f_{ji}}{\partial u_k}\frac{\partial u_k}{\partial x_i}}_{0 \text{ by PDE}}\right) \quad \text{(here we changed the } l \text{ index to } j\text{)} \\
&= 0.
\end{aligned}
$$

We observe that satisfying this inequality implies [1.27] is also satisfied. □

Remark 1.30. The same result holds in the weak sense if u is piecewise smooth and continuous.

Entropy is one way to consider physically relevant solutions; another method is to consider (as in the scalar case) *viscous regularization*. For $\epsilon > 0$, we consider the problem:

$$
\partial_t u_\epsilon + \nabla f(u_\epsilon) - \epsilon\Delta u_\epsilon = 0; \quad u_\epsilon(x,0) = u_0(x).
$$

We further say that u is a *vanishing viscosity solution* to [1.3.2] if $\|u_\epsilon - u\|_{L^1} \to 0$ as $\epsilon \to 0$.

Remark 1.31 (Lack of Uniqueness)**.** Unlike in the scalar case, entropy solutions and viscosity solutions do *not* guarantee uniqueness. Indeed, it has been shown that there exists infinitely many weak solutions that also satisfy the entropy inequalities.

Lastly, we consider a few examples

Example 1.32 (Linear Wave Equation)**.** Consider the system:

$$\begin{cases} \partial_t u + \nabla \cdot \mathbf{v} = 0 \\ \partial_t \mathbf{v} + c^2 \nabla u = 0 \end{cases}. \tag{1.29}$$

where $u \in \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^d$.

First we take the time derivative of the first equation:

$$\partial_t \left[\partial_t u + \nabla \cdot \mathbf{v}\right] = \partial_{tt} u + \partial_t \nabla \cdot \mathbf{v} = 0.$$

Then we take the divergence of the second equation:

$$\nabla \cdot \left[\partial_t \mathbf{v} + c^2 \nabla u\right] = \nabla \cdot \partial_t \mathbf{v} + c^2 \nabla \cdot \nabla u = \nabla \cdot \partial_t \mathbf{v} + c^2 \Delta u = 0.$$

With appropriate assumptions on $\mathbf{v}$–namely that it has continuous second partial derivatives–we have:

$$\partial_t \nabla \cdot \mathbf{v} = \partial_t \left(\sum_{i \in \{1:d\}} \partial_{xi} v_i \right) = \sum_{i \in \{1:d\}} \partial_t \partial_{xi} v_i = \sum_{i \in \{1:d\}} \partial_{xi} \partial_t v_i = \nabla \cdot (\partial_t \mathbf{v}).$$

Then, by subtracting the second from the first, we obtain the linear wave equation:

$$\partial_{tt} u - c^2 \Delta u = 0. \tag{1.30}$$

We can recast this as a conservation problem: let $\mathbf{u} = \begin{pmatrix} u \\ \mathbf{v}^T \end{pmatrix}$ (and thus $\mathbf{u} \in \mathbb{R}^{1+d}$) and let

$$\mathbf{f}(u) := \begin{pmatrix} \mathbf{v}^T \\ c^2 u \mathbf{I}_d \end{pmatrix} = \begin{bmatrix} v_1 & v_2 & \dots & v_d \\ c^2 u & 0 & \dots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \dots & c^2 u \end{bmatrix}.$$

Then, $\partial_t \mathbf{u} = \begin{pmatrix} \partial_t u \\ \partial_t \mathbf{v} \end{pmatrix}$ and:

$$\nabla \cdot \mathbf{f}(u) = \sum_{i \in \{1:d\}} \partial_{xi} \begin{bmatrix} f_{1i}(u) \\ \vdots \\ f_{d+1,i}(u) \end{bmatrix} = \sum_{i \in \{1:d\}} \begin{bmatrix} \partial_{xi} v_1 \\ 0 \\ \vdots \\ 0 \\ c^2 \partial_{xi} u \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Thus, the system is:

$$\partial_t \mathbf{u} + \nabla \cdot f(u) = \begin{bmatrix} \partial_t u \\ \partial_t v_1 \\ \vdots \\ \partial_t v_d \end{bmatrix} + \sum_{i \in \{1:d\}} \begin{bmatrix} \partial_{xi} v_i \\ \vdots \\ c^2 \partial_{xi} u \end{bmatrix} = \begin{bmatrix} \partial_t u + \sum_{i \in \{1:d\}} \partial_{xi} v_i \\ \partial_t v_1 + c^2 \partial_{x1} u \\ \vdots \\ \partial_t v_d + c^2 \partial_{xd} u \end{bmatrix}.$$

The first row of this matrix gives the first equation of the original system, whereas the rest of the matrix gives the system for the second equation. Finally, we give some computations. Let $\mathbf{n}$ be a unit vector in $\mathbb{R}^d$. Then:

$$\mathbf{f}(\mathbf{u})\cdot\mathbf{n} = \begin{bmatrix} v_1n_1 + \cdots + v_dn_d \\ c^2un_1 + 0 + \cdots + 0 \\ \vdots \\ 0 + \cdots + 0 + c^2un_d \end{bmatrix} = \begin{bmatrix} \mathbf{n}\cdots\mathbf{n} \\ c^2un_1 \\ \vdots \\ c^2n_d \end{bmatrix} = \begin{bmatrix} \mathbf{v}\cdot\mathbf{n} \\ c^2u\mathbf{n} \end{bmatrix}.$$

Lastly:

$$D(\mathbf{f}(\mathbf{u})\cdot\mathbf{n}) = D\begin{pmatrix} \mathbf{v}\cdot\mathbf{n} \\ c^2un_1 \\ \vdots \\ c^2un_d \end{pmatrix} = \begin{bmatrix} 0 & n_1 & n^2 & \dots & n_d \\ c^1n_1 & 0 & \dots & & 0 \\ \vdots & & & & \vdots \\ c^2n_d & 0 & \dots & & 0 \end{bmatrix}.$$

It can be shown that there are $(d+1)$ eigenpairs of the above matrix, and that they are $(c, (1, c\mathbf{n}^T)^T)$ and $(0, (0, v_l^T)^T)$ for all $l \in \{1 : d-1\}$. Furthermore, the vectors $\{v_l\}_{l\in\{1:d-1\}}$ are such that $\{\mathbf{n}, \mathbf{v}_1, \dots, \mathbf{v}_{d-1}\}$ form an o.n. basis for $\mathbb{R}^d$.

Example 1.33 (p-systems)**.** Consider the model for the 1D isentropic gas:

$$\begin{cases} \partial_t v - \partial_x u = 0 \\ \partial_t u + \mathrm{d}xp(v) = 0. \end{cases} \tag{1.31}$$

Here, $d = 1,\ m = 2$ and the map $v \mapsto p(v)$ is the pressure, u =velocity, v =specific volume. We require that $0 < p'', p' < 0$. To make this problem a conservation equation, we define:

$$\mathbf{u} = \begin{bmatrix} v \\ u \end{bmatrix},\ n = \pm e_x,\ \mathbf{f}(\mathbf{u}) = \begin{pmatrix} -u \\ p(v) \end{pmatrix}.$$

Therefore, we compute:

$$\partial_x\mathbf{f}(\mathbf{u}) = \begin{pmatrix} -\partial_x u \\ p'(v)\partial_x v \end{pmatrix} = \underbrace{\begin{bmatrix} 0 & -1 \\ p'(v) & 0 \end{bmatrix}}_{\mathbf{A}(\mathbf{u})}\underbrace{\begin{bmatrix} \partial_x v \\ \partial_x u \end{bmatrix}}_{\partial_x\mathbf{u}}.$$

Thus,

$$\mathbf{A}(n, v) = \sum_{l\in\{1:d\}} n_l\mathbf{A}_l(\mathbf{v}) = n\mathbf{A}(\mathbf{v}),\ n = \pm 1.$$

Following the definition of hyperbolicity, the system is hyperbolic if $\mathbf{A}(n, \mathbf{v})$ has real eigenvalues and is diagonalizable. Computation yields:

$$\det(\mathbf{A}(n, \mathbf{v}) - \lambda I) = 0 \implies \lambda^2 - p'(v) = 0 \implies \lambda = \pm\sqrt{-p'(v)}.$$

(Quickly observe that the above computation relies on $n = \pm 1$.) Then, solving the system:

$$\begin{bmatrix} 0 & -1 \\ p'(v) & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \sqrt{-p'(v)}x_1 \\ \sqrt{-p'(v)}x_2 \end{bmatrix}.$$

Gives $-x_2 = \sqrt{-p'(v)}x_1$, which implies that $\begin{pmatrix} 1 \\ -\sqrt{-p'(v)} \end{pmatrix}$ is a an eigenvector. Likewise, $-x_2 = -\sqrt{-p'(v)}x_1$ implies that $\begin{pmatrix} 1 \\ \sqrt{-p'(v)} \end{pmatrix}$ is an eigenvector. And so the system is hyperbolic.

1.3.4 Riemann Problem

Consider the first order quasi-linear equation:

$$\partial_t \mathbf{u} + \nabla \cdot \mathbf{f}(u) = 0. \tag{1.32}$$

Now we consider a 1D Riemann problem from [1.32]: Let $\mathbf{n} \in \mathbb{R}^d$ be a unit vector. We seek $\mathbf{v} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}^m$ such that:

$$\partial_t \mathbf{v} + \underbrace{\partial_x F(u)}_{\partial_x(\mathbf{f}(u)\cdot\mathbf{n}) \text{ can be viewed as } \mathbf{A}(u,\mathbf{n})\cdot\partial_x u} = 0; \quad \mathbf{v}(x,0) = \begin{cases} v_R & x \le 0 \\ v_L & 0 < x \end{cases} \quad F(u) := \mathbf{f}(u)\cdot n. \tag{1.33}$$

(Recall that $\mathbf{A}(\mathbf{n},\mathbf{v}) = \sum_{l\in\{1:d\}} \mathbf{A}_l(\mathbf{v}) n_l$, so that $(\mathbf{A}(\mathbf{n},\mathbf{v}) = \sum_{l\in\{1:d\}} \frac{\partial f_{il}}{\partial v_j}$ for $i,j \in \{1:m\}$, so that $\mathbf{A}(\mathbf{n},\mathbf{v})$ is a square matrix).

Let $(\lambda_l(\mathbf{v}), \mathbf{r_l}(\mathbf{v})$ be an eigenpair of $\mathbf{A}(\mathbf{n},\mathbf{v})$, (so $\mathbf{A}(\mathbf{n},\mathbf{v}) r_l(\mathbf{v}) = \lambda_l(\mathbf{v}) r_l(\mathbf{v})$. We adopt the convection that $\lambda_1(\mathbf{v}) \le \cdots \le \lambda_m(\mathbf{v})$. We assume that the map $\mathbf{v} \mapsto (\lambda_l(\mathbf{v}), r_l(\mathbf{v}))$ is C^1.

Definition 1.34. Let $l \in \{1:m\}$, $\mathbf{n} \in \mathbb{R}^d$ be a unit vector. The l-th eigenpair is *genuinely nonlinear* if $D\lambda_l(\mathbf{v}) \cdot r_l(\mathbf{v}) = 0$ for all $\mathbf{v} \in \mathcal{A}$, and is *linearly degenerate* if $D\lambda_l(\mathbf{v}) \cdot r_l(\mathbf{v}) = 0$ for all $\mathbf{v} \in \mathcal{A}$.

Example 1.35 (Scalar Case: $d = m = 1$.)**.**

$$\partial_t u + \partial_x(f(u)) = \partial_t u + f'(u)\partial_x u = 0.$$

Then $\mathbf{A}(\mathbf{v},\mathbf{n}) = \sum_{l\in\{1:d\}} n_l \mathbf{A}_l(\mathbf{v}) = 1 f'(u) = f'(u)$. Thus, $(f'(u), e_x)$ is the only eigenpair. Thus, the system is genuinely nonlinear iff $D\lambda(u) = Df'(u) = f''(u) \neq 0$ and is linearly degenerate iff $f''(u) = 0$.

Moving forward, we consider the problem [1.33], but for simplicity that all of the eigenpairs of $\mathbf{A}(\mathbf{n},\mathbf{v})$ are either genuinely nonlinear or linearly degenerate. Furthermore, we make the following normalizations: if $r_l(\mathbf{v})$ is linearly degenerate, then set $\|r_l(\mathbf{v})\| = 1$, and if $r_l(\mathbf{v})$ is genuinely nonlinear, then set $D\lambda_l(\mathbf{v}0 \cdot r_l(\mathbf{v}) = 1$.

As in the scalar, we would like to find a self-similar solution to [1.33]. That is, we seek $\mathbf{w}(x/t)$ such that $\mathbf{u}(x,t) = \mathbf{w}(x/t)$. Let us denote $\xi := x/t$.

Lemma 1.36. *The map $\mathbf{w}(\xi)$ is a self-similar weak solution to [1.33] iff $\mathbf{A}(\mathbf{w}(\xi),\mathbf{n})\mathbf{w}'(\xi) = \xi\mathbf{w}(\xi)$.*

Proof. Suppose $\mathbf{w}(\xi)$ is a solution. Recall:

$$\mathbf{A}(\mathbf{w},\mathbf{n}) = \begin{bmatrix} \sum_{l\in\{1:d\}} n_l \frac{\partial f_{1l}}{\partial w_1} & \cdots & \sum_{l\in\{1:d\}} n_l \frac{\partial f_{1l}(w)}{\partial w_m} \\ \vdots & & \vdots \\ \sum_{l\in\{1:d\}} n_l \frac{\partial f_{ml}(w)}{\partial w_1} & \cdots & \sum_{l\in\{1:d\}} \frac{\partial f_{ml}}{\partial w_m} \end{bmatrix}.$$

Next, we compute:

$$\partial_x \mathbf{w}\left(\frac{x}{t}\right) = \begin{bmatrix} \partial_x w_1(x/t) \\ \vdots \\ \partial_x w_m(x/t) \end{bmatrix} = \begin{bmatrix} w_1'(x/t)\frac{1}{t} \\ \vdots \\ w_m'(x/t)\frac{1}{t} \end{bmatrix} = \begin{bmatrix} w_1'(x/t) \\ \vdots \\ w_m'(x/t) \end{bmatrix} \frac{1}{t} = \mathbf{w}'\left(\frac{x}{t}\right)\frac{1}{t}.$$

Therefore:

$$\partial_x(\mathbf{f}(\mathbf{w})\cdot\mathbf{n}) = \begin{bmatrix} \sum_{l\in\{1:d\}} n_l \frac{\partial f_{1l}(\mathbf{w})}{\partial w_1}\frac{\partial w_1}{\partial x}\frac{1}{t} + \cdots + \sum_{l\in\{1:d\}} n_l \frac{f_{1l}(\mathbf{w})}{w_m}\frac{\partial w_m}{\partial x}\frac{1}{t} \\ \vdots \\ \sum_{l\in\{1:d\}} n_l \frac{\partial f_{ml}(\mathbf{w})}{\partial w_l}\frac{\partial w_1}{\partial x}\frac{1}{t} + \cdots + \sum_{l\in\{1:d\}} n_l \frac{\partial f_{ml}(\mathbf{w})}{\partial w_m}\frac{\partial w_m}{\partial x}\frac{1}{t} \end{bmatrix} = \frac{1}{t}\mathbf{A}(\mathbf{w},\mathbf{n})\mathbf{w}'(x/t).$$

Lastly:

$$\partial_t \mathbf{w}\left(\frac{x}{t}\right) = -\frac{x}{t^2}\mathbf{w}'\left(\frac{x}{t}\right).$$

Thus:

$$\partial_t \mathbf{w}\left(\frac{x}{t}\right) + \partial_x\left(\mathbf{f}(\mathbf{w})\cdot\mathbf{n}\right) = -\frac{x}{t^2}\mathbf{w}'\left(\frac{x}{t}\right) + \mathbf{A}(\mathbf{w},\mathbf{n})\mathbf{w}'\left(\frac{x}{t}\right)\frac{1}{t} = 0.$$

Therefore:

$$\mathbf{A}(\mathbf{w},\mathbf{n})\mathbf{w}'\left(\frac{x}{t}\right) = \xi\mathbf{w}'(\xi).$$

The converse follows by a reverse argument.

□

The above lemma is only possible if $\mathbf{w}(\xi)$ is constant or $(\xi, \mathbf{w}(\xi))$ is an eigenpair of $\mathbf{A}(\mathbf{w},\mathbf{n})$. If is is an eigenpair, then there exists an $l \in \{1:d\}$ such that $\lambda_l(\mathbf{w}(\xi)) = \xi$ and $\mathbf{w}'(\xi)$ is proportional to $r_l(\mathbf{w}(\xi))$, ie, there exist some $\gamma(\xi)$ such that $\mathbf{w}(\xi) = \gamma(\xi) r_l(\mathbf{w}(\xi))$.

Lemma 1.37. *If $r_l(\mathbf{w}(\xi))$ is genuinely nonlinear and $\lambda_l(\mathbf{w}(\xi)) = \xi$ and $w'(\xi)$ is proportional to $r_l(\mathbf{w}(\xi))$, then $\mathbf{w}'(\xi) = r_l(\mathbf{w}(\xi))$.*

Proof. Suppose $(\lambda_l(\mathbf{w}(\xi)), r_l(\mathbf{w}(\xi)))$ is genuinely nonlinear. Then:

$$\frac{d}{d\xi}\lambda_l(\mathbf{w}(\xi)) = \frac{d}{d\xi}\xi \implies D\lambda_l(\mathbf{w}(\xi))\cdot\mathbf{w}'(\xi) = 1 \implies \gamma(\xi)\underbrace{D\lambda_l(\mathbf{w}(\xi))r_l(\xi))}_{1} = 1 \implies \mathbf{w}'(\xi) = r_l(\mathbf{w}(\xi)).$$

□

Lemma 1.38. *Suppose that the l-th eigenpair is genuinely nonlinear. Let $\mathbf{u}_z \in \mathcal{A}$ (that is, pick a state), $\xi_z = \lambda_l(u_z)$. Let $\delta > 0$ be such that $\mathbf{w} \in C^1((\xi_z - \delta, \xi_z + \delta), \mathbb{R}^m)$ solves the ODE $\mathbf{w}'(\xi) = r_l(\mathbf{w}(\xi))$ with $(w)(\xi_z) = \mathbf{u}_z$. Let $I := (\xi_z - \delta, \xi_z + \delta)$*

1. *The identity $\lambda_l(\mathbf{w}(\xi)) = \xi$ holds for all $\xi \in I$.*
2. *Let $\xi_L \in (\xi_z - \delta, \xi_z)$ and set $\mathbf{u}_L := \mathbf{w}(\xi_L)$ (ie, pick a left state). Let $\xi_R \in (\xi_z, \xi_z + \delta)$, $\mathbf{u}_R = \mathbf{w}(\xi_R)$ (ie, pick a right state). Then, $\lambda_l(\mathbf{u}_L) < \lambda_l(\mathbf{u}_R)$ and the function :*

$$u(x,t) := \begin{cases} \mathbf{u}_L & x/t \le \lambda_l(\mathbf{u}_L) \\ \mathbf{w}(\xi) & \lambda_l(\mathbf{u}_L) < x/t < \lambda_l(\mathbf{u}_R) \\ \mathbf{u}_R & \lambda_l(\mathbf{u}_R) < x/t \end{cases} \tag{1.34}$$

is a self-similar weak solution to the 1D Riemann problem.

Proof. Since we normalized r_l, we have:

$$\frac{d}{d\xi}\left(\xi - \lambda_l(\mathbf{w}(\xi))\right) = 1 - D\lambda_l(\mathbf{w}(\xi))\mathbf{w}'(\xi) = 1 - 1 = 0.$$

Thus, $\xi - \lambda_l(\mathbf{w}(\xi))$ is constant for all $\xi \in I$, and since $\xi_z - \lambda_l(\mathbf{w}(\xi_z)) = 0$, $\xi = \lambda_l(\mathbf{w}(\xi))$ for all $\xi \in I$.

Since $\xi_L \in (\xi_z - \delta, \xi_z)$, (1) gives that $\lambda_L = \lambda_l(\mathbf{w}(\xi_L)) = \lambda_l(\mathbf{u}_L)$. Likewise, $\xi_R = \lambda_l(\mathbf{u}_R)$. Since $\xi_L < \xi_R$, we conclude that $\lambda_l(\mathbf{u}_L) < \lambda_l(\mathbf{u}_R)$. Note that this give that u as defined above is continuous. Thus, u is piecewise smooth, and so we only need to satisfy the PDE on the three areas: $x/t < \lambda_l(\mathbf{u}_L)$, $\lambda_l(\mathbf{u}_L) < x/t < \lambda_l(\mathbf{u}_R)$ and $\lambda_l(\mathbf{u}_R) < x/t$. In the first and third regions, u is constant, and thus satisfies the PDE. In the middle region, we observe that $\mathbf{w}'(\xi) = r_l(\mathbf{w}(\xi))$, so the condition for the lemma holds and is solution on the interval. □

Lemma 1.39 (Rankine Hugoniot)**.** *Let $s \in \mathbb{R}_+$. The function:*

$$u(x,t) = \begin{cases} \mathbf{u}_L & x/t \leq s \\ \mathbf{u}_R & s < x/t \end{cases}$$

is a weak solution to the 1D problem iff s is such that

$$\mathbf{f}(\mathbf{u}_L) \cdot n - \mathbf{f}(\mathbf{u}_R) \cdot n = s(\mathbf{u}_L - \mathbf{u}_R). \tag{1.35}$$

Proof. Let $t \leq \frac{1}{s}$. We integrate:

$$\begin{aligned}
\int_{-1}^{1}\int_{0}^{t} \partial_t u d\tau \mathrm{d}x &= \int_{-1}^{1} u(x,t) - u(x,0)\mathrm{d}x \\
&= \int_{-1}^{1} u(x,t)\mathrm{d}x - [\mathbf{u}_R + \mathbf{u}_L] \\
&= \int_{-1}^{ts} u(x,t)\mathrm{d}x + \int_{ts}^{1} u(x,t)\mathrm{d}x - [\mathbf{u}_R + \mathbf{u}_L] \\
&= \int_{-1}^{ts} \mathbf{u}_L \mathrm{d}x + \int_{ts}^{1} \mathbf{u}_R \mathrm{d}x - [\mathbf{u}_R - \mathbf{u}_L] \\
&= \mathbf{u}_L[ts+1] + \mathbf{u}_R[1-ts] - [\mathbf{u}_R + \mathbf{u}_L] \\
&= \mathbf{u}_L ts - \mathbf{u}_R ts = ts[\mathbf{u}_L - \mathbf{u}_R]
\end{aligned}$$

$$\begin{aligned}
\int_0^1 \int_{-1}^{1} \partial_x(\mathbf{f}(u)\cdot\mathbf{n})\mathrm{d}x d\tau &= \int_0^t \mathbf{f}(u(1,\tau))\cdot\mathbf{n} - bff(u(-1,\tau))\cdot\mathbf{n}\ d\tau \\
&= \int_0^t \mathbf{f}(u_R)\cdot\mathbf{n} - \mathbf{f}(u_L)\cdot n d\tau \\
&= t[\mathbf{f}(u_R)\cdot n - \mathbf{f}(u_L)\cdot n].
\end{aligned}$$

Thus:

$$s(u_L - u_R) = \mathbf{f}(u_L)\cdot\mathbf{n} - \mathbf{f}(u_R)\cdot\mathbf{n}.$$

□

Lemma 1.40 (Contact Discontinuity)**.** *Assume that the l-th eigenpair is linearly degenerate. let $u_z \in \mathcal{A}$ and let $\xi_z = \lambda(u_z)$. Let $\delta > 0$ be such that $z \in C^1((\xi_z - \delta, \xi_z + \delta))$ solves the ODE $z'(\xi) = r_l(z(\xi))$ with $z(\xi_z) = u_z$.*

1. *The identity $\lambda_l(z(\xi)) = \lambda_l(u_z)$ holds for all $\xi \in I := (\xi_z - \delta, \xi_z + \delta)$.*
2. *Let $\xi_L \in (\xi_L - \delta, \xi_z)$ and let $u_L := z(\xi_L)$. Let $\xi_R \in (\xi_z, \xi_R + \delta)$, $u_R := z(\xi_R)$. Then the function defined by:*

$$u(x,t) = \begin{cases} u_L & x/t < \lambda_l(u_z) \\ u_R & x/t > \lambda_l(u_z). \end{cases} \tag{1.36}$$

is a self-similar weak solution to the Riemann Problem.

Proof. We calculate:

$$\frac{d}{d\xi}\lambda_l(z(\xi)) = D\lambda_l(z(\xi))\cdot z'(\xi) = D\lambda_l(z(\xi))\cdot r_l(z(\xi)) = 0.$$

Thus, $\lambda_l(z(\xi))$ is constant, so $\lambda_l(z(\xi)) = \lambda_l(u_L)$ for all $\xi \in I$.

$$\begin{aligned}
f(u_R)\cdot n - f(u_L)\cdot n &= \int_{\xi_L}^{\xi_R} \frac{d}{d\xi} f(z(\xi))\cdot n d\xi \\
&= \int_{\xi_L}^{\xi_R} Df(z(\xi))\cdot n \cdot z'(\xi) d\xi \\
&= \int_{\xi_L}^{\xi_R} \mathbf{A}(z(\xi), n)\cdot r_l(z(\xi)) d\xi \\
&= \int_{\xi_L}^{\xi_R} \lambda_l(z(\xi)) r_l(z(\xi)) d\xi \\
&= \int_{\xi_L}^{\xi_R} \lambda_l(u_z) z'(\xi) d\xi \\
&= \lambda_l(u_z)\int_{\xi_L}^{\xi_R} z'(\xi) d\xi \\
&= \lambda_l(u_z)[z(\xi_R) - z(\xi_L)] \\
&= \lambda_l(u_z)(u_R - u_L).
\end{aligned}$$

Thus by Rankine-Huguinot we have a solution. □

Note 1.41. Contact discontinuities are discontinuities that simply travel in time—compare with shocks that come from interacting velocities.

Lemma 1.42 (Shock)**.** *Let $u_z \in \mathcal{A}$, assume that the eigenvalue $\lambda_l(u_z)$ has multiplicity 1 and let $\xi = \lambda_1(u_z)$.*

1. *There exists a $\delta > 0$ and functions $s_l \in C^0((\xi_z - \delta, \xi_z + \delta), \mathbb{R})$, $z_l \in C^0((\xi_z - \delta, \xi_z + \delta), \mathbb{R}^m)$ such that for all $\xi \in I$*
$$(\mathbf{f}(z_l(\xi)) - \mathbf{u_z})\cdot n = s_l(\xi)(z_l(\xi) - u_z).$$
2. *Fix $\xi \in I$. Set $s = s_1(\xi)$. If $\xi < \xi_z$, set $u_l := z_l(\xi)$ and $u_R := u_z$, whereas if $\xi_z < \xi$, set $u_L := u_z$, and $u_R := z_l(\xi)$. Then the function defined by:*
$$u(x,t) = \begin{cases} u_L & x/t \le s \\ u_R & s < x/t \end{cases}$$

is a self-similar solution to the Riemann Problem.

Note 1.43. Shocks are the result of collisions/ velocities hitting each other.

Theorem 1.44. *Assume that the Riemann problem is strictly hyperbolic and that for all $l \in \{1 : m\}$ the l-th eigenpair is either genuinely nonlinear or linearly degenerate. Then there exists a $\delta > 0$ such that for every pair $(u_L, u_R) \in \mathcal{A}^2$ such that $\|u_R - u_L\| \le \delta$, the Riemann problem has a weak solution that consists of at most $m + 1$ constant states separated by expansion waves, shocks, or contact discontinuities, and this solution is a vanishing-viscosity solution.*

Note 1.45. That is, the solutions to Riemann problems can be stitched together—"a miracle"-JLG.

This theorem implies that there are at most $2m$ numbers $\{\lambda_l^{\pm}\}_{l\in\{1:m\}}$ such that $\lambda_1^- \le \lambda_1^+ \le \cdots \le \lambda_m^- \le \lambda_m^+$ which determine up to $2m + 1$ areas in the $x - t$ plane such that the solution is u_L in the first sector $\{x/t \le \lambda_1^-\}$, u_R in the last $\{x/t \ge \lambda_m^+\}$, and in between is either constant or an expansion wave. If $\lambda_l^- = \lambda_l^+$, then we have a shock or a contact discontinuity. This gives rise to the *Riemann Fan*.

1.3.5 Maximum speed and averages

Throughout this section, let $\mathbf{n} \in \mathbb{R}^d$ be a unit vector. Let $\lambda_{\max} \geq \max(|\lambda_1^-(\mathbf{n}, u_L, u_R)|, |\lambda_m^+(\mathbf{n}, u_L, u_R)|)$.

Definition 1.46 (Max Wave Speed)**.** Let $(u_R, u_L) \in \mathcal{A}^2$ and $\mathbf{n} \in \mathbb{R}^d$ be a unit vector. The number

$\max(|\lambda_1^-(u_L, u_R)|, |\lambda_m^+(u_L, u_R)|)$ get a max on wave speed, either shock speed or from expansion waves

is called the maximum wave speed in the Riemann problem. Any read numbers $\lambda_{\max}(u_L, u_R)$ satisfying the bound above is called an *upper bound on the maximum wave speed* in the Riemann problem.

Let $u(u_R, u_L)(x, t)$ be the vanishing viscosity solution in the above theorem. Observe that $x/t \geq \lambda(u_L, u_R) \geq \lambda_m^+(u_L, u_R)$ and $x/t \leq -\lambda_{\max}(u_L, u_R) \leq \lambda_1^-$ implies that:

$$(\text{Solution to 1D Riemann Problem}) \leftarrow u(u_L, u_R)(x, t) = \begin{cases} u_L & x \leq -t\lambda_{\max}(u_L, u_R) \\ u_R & x \geq t\lambda_{\max}(u_L, u_R). \end{cases}.$$

Let $0 \leq t\lambda_{max}(u_L, u_R) \leq 1/2$; we define the *Riemann Average* to be:

$$\overline{u}(t, u_L, u_R) := \int_{-1/2}^{1/2} u(u_L, u_R)(x, t)\mathrm{d}x. \tag{1.37}$$

We observe that if $(\eta, \mathbf{q})$ is an entropy pair for the quasi linear system, then $(\eta, \mathbf{q} \cdot \mathbf{n})$ is an entropy pair for the Riemann problem.

Lemma 1.47. *Recall that $\overline{u}(t, u_L, u_R)$ is defined for $0 \leq t\lambda_{\max}(u_L, u_R) \leq 1/2$. Let $(\eta, \mathbf{q})$ be an entropy pair for the quasi-linear system. Then:*

$$\overline{u}(t, u_L, u_R) = \frac{1}{2}(u_L + u_R) - f(\mathbf{f}(u_R) \cdot \mathbf{n} - \mathbf{f}(u_L) \cdot \mathbf{n}) \tag{1.38}$$

$$\eta(\overline{u}(t, u_L, u_R)) \leq \frac{1}{2}(\eta(u_L) + \eta(u_R)) - t(\mathbf{q}(u_R) \cdot \mathbf{n} - \mathbf{q}(u_L) \cdot \mathbf{n}). \tag{1.39}$$

Proof. We integrate:

$$\int_{-1/2}^{1/2} \int_0^t \partial_t u + \partial_x(\mathbf{f}(u) \cdot n)\mathrm{d}\tau\mathrm{d}x = 0.$$

$$\int_{-1/2}^{1/2} \int_0^t \partial_t u d\tau\mathrm{d}x = u(x, t) - u(x, 0) \implies \int_{-1/2}^{1/2} u(x, t) - u(x, 0)\mathrm{d}x = \int_{-1/2}^{1/2} u(x, t)\mathrm{d}x - \frac{1}{2}(u_L + u_R).$$

$$\int_0^t \int_{-1/2}^{1/2} \partial_x(\mathbf{f}(u) \cdot \mathbf{n})\mathrm{d}x d\tau = \int_0^t \mathbf{f}(u(1/2, \tau)) \cdot \mathbf{n} - \mathbf{f}(u(-1/2, \tau)) \cdot \mathbf{n} d\tau.$$

The inequalities $\frac{-1}{2\tau} \leq \frac{-1}{2t} \leq -\lambda_{\max}(u_L, u_R)$ and $\frac{1}{2\tau} \geq \frac{1}{2t} \geq \lambda_{\max}(u_L, u_R)$ imply that $u(-1/2, \tau) = u_L$ and $u(1/2, \tau) = u_R$. Thus:

$$\int_0^t \int_{-1/2}^{1/2} \partial_x(\mathbf{f}(u) \cdot \mathbf{n})\mathrm{d}x d\tau = \int_0^t f(u_R) \cdot \mathbf{n} - f(u_L) \cdot n d\tau = t(\mathbf{f}(u_R) \cdot \mathbf{n} - \mathbf{f}(u_L) \cdot \mathbf{n}).$$

Therefore:

$$\int_0^t \int_{-1/2}^{1/2} \partial_t u - \partial_x(\mathbf{f}(u) \cdot \mathbf{n})\mathrm{d}x d\tau = 0 \implies \overline{u}(t, u_L, u_R) = \int_{-1/2}^{1/2} u(x, t)\mathrm{d}x = \frac{1}{2}(u_L + u_R) - t(\mathbf{f}(u_R) \cdot \mathbf{n} - \mathbf{f}(u_R) \cdot \mathbf{n}).$$

A similar argument shows that $\int_0^t \int_{-1/2}^{1/2} \partial_t \eta(u) + \partial_x \mathbf{q}(u)\mathrm{d}x d\tau \leq 0$ implies that

$$\int_{-1/2}^{1/2} \eta(u(x, t))\mathrm{d}x \leq \frac{1}{2}(\eta(u_L) + \eta(u_R)) - t(\mathbf{q}(u_R) - \mathbf{q}(u_L)).$$

Jenssen's Inequality gives the result. □

1.3.6 Invariant Sets

Invariant sets replace the notion of maximum principles for hyperbolic systems.

Definition 1.48. A convex set $\mathcal{B} \subset \mathcal{A} \subset \mathbb{R}^m$ is said to be *invariant* for the hyperbolic system if for every $(u_L, u_R) \in \mathcal{B}$ and every $\mathbf{n} \in \mathbb{R}^d$, unit vector, the vanishing viscotiy solution to the Riemann problem is in $\overline{\mathcal{B}}$ for a.e. $x \in \mathbb{R}$ and a.e $t > 0$ with $t\lambda_{\max}(u_L, u_R) \leq 1/2$.

Lemma 1.49 (Riemann Average)**.** *Let $\mathcal{B} \subset \mathcal{A} \subset \mathbb{R}^m$ be an invariant set. Let $(u_R, u_L) \in \mathcal{B}^2$ and let $\mathbf{n} \in \mathbb{R}^d$ be a unit vector.*

1. *If $t\lambda_{\max}(u_L, u_R) \leq 1/2$, the $\overline{u}(t, u_L, u_R) \in \mathcal{B}$.*
2. *If $t\lambda_{\max}(u_L, u_R) < 1/2$ and $(u_L, u_R) \in int(\mathcal{B})$, then $\overline{u}(t, u_L, u_R) \in int(\mathcal{B})$.*

Proof. The function $d(v) := \inf_{z \in \mathcal{B}} \|v - z\|_{l^2}$ is convex because $\mathcal{B}$ is convex. Then, by Jenssen's Inequality and the fact that $u(u_L, u_R)(x, t) \in \mathcal{B}$, we have:

$$d(\overline{u}(t, u_L, u_R)) = d\left(\int_{-1/2}^{1/2} u(x,t)\mathrm{d}x\right) \leq \int_{-1/2}^{1/2} d(u(u_L, u_R)(x,t))\mathrm{d}x = 0.$$

Thus, by convexity, $\overline{u}(t, u_L, u_R) \in \mathcal{B}$.
For (2), define

$$w(t) := \frac{1}{2\lambda_{max}} \int_{-\lambda_{max}t}^{\lambda_{\max}t} u(u_L, u_R)(x,t)\mathrm{d}x.$$

Thus, $\overline{u}(t, u_L, u_R)(x,t) = (1 - 2\lambda_{\max}t) - \frac{1}{2}(u_L + u_R) + 2\lambda_{\max}tw(t)$. The above argument shows $w(t) \in \overline{\mathcal{B}}$. Since $1/2(u_L + u_R) \in int(\mathcal{B})$, $\overline{u}$ cannot be on the boundary and thus is in the interior. □

1.4 Chapter 81: First Order Approximation

1.4.1 Scalar Conservation Equations

Naturally, we being our approximation schemes with scalar-valued conservation laws. (For simplicity, we shall enforce Dirichlet boundary conditions.)

Finite element space

We will use finite elements in our space discretization and forward Euler for our time discretization.

- Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape regular family of matching meshes, ie, the meshes cover D exactly.
- $(\hat{K}, \hat{P}, \hat{\Sigma})$ be the finite element.
- Let $T_k : \hat{K} \to K$ be the geometric mapping for $K \in \mathcal{T}_h$.
- The finite element space:

$$P_k^g(\mathcal{T}_h) = \{v \in C^0(\overline{D}, \mathbb{R}) : v|_K \circ T_k \in \hat{P}, \forall K \in \mathcal{T}_h\}.$$

- Reference shape functions: $\{\hat{\theta}_i\}_{i \in \mathcal{N}}$ with $\mathcal{N} := \{1 : n_{sh}\}$. We also require partition of unity, ie, $\sum_{i \in \mathcal{N}} \hat{\theta}_i(x) = 1$ for all $x \in \hat{K}$.
- Global shape functions for $P_K^\theta(\mathcal{T}_h)$: $\{\phi_i\}_{i \in \mathcal{A}_h}$ with $\mathcal{A}_h := \{1 : I\}$, $I := dim(P_k^g(\mathcal{T}_h))$. Then, let $j_d of : \mathcal{T}_h \times \mathcal{N} \mapsto \mathcal{A}_h$ be the connectivity array such that $\phi_{j_d of(k,i)}|_K = \hat{\theta}_i \circ T_k^{-1}$ for all $(k, i) \in \mathcal{T}_h \times \mathcal{N}$. Thus, we observe that $\sum_{i \in \mathcal{A}_h} \phi_i(x) = 1$ for all $x \in \overline{D}$.
- Further recall that $\mathcal{A}_h$ can be partitioned: $\mathcal{A}_h = \mathcal{A}_h^o \cup \mathcal{A}_h^\partial$.

- We now define $I(i) := \{j \in \mathcal{A}_h : \ \phi_i \phi_j \neq 0\}$. (This is also called the *stencil*)
- Readily observe that $j \in I(i)$ iff $i \in I(j)$.
- We define the mass matrix M by $M_{ij} = \int_D \phi_i \phi_j \mathrm{d}x$.
- The *lumped mass matrix* $\overline{M}$ is defined as the diagonal matrix $\overline{M}_i = \int_D \phi_i(x)\mathrm{d}x$. Observe that the partition of unity implies that:

$$\int_D \phi_i(x)\mathrm{d}x = \int_D 1\phi_i(x)\mathrm{d}x = \int_D \sum_{j\in\mathcal{A}_h} \phi_j(x)\phi_i(x)\mathrm{d}x = \sum_{j\in\mathcal{A}_h} \int_D \phi_i(x)\phi_j(x)\mathrm{d}x = \sum_{j\in\mathcal{A}_h} m_{ij}.$$

  We also make a key assumption: $m_i > 0$ for all $i \in \mathcal{A}_h$.

The Scheme

First, we let $u_h^0 = \sum_{i\in\mathcal{A}_h} U_i^0 \phi_i \in P_k^g(\mathcal{T}_h)$ be a "reasonable" approximation of u_0. We let t_n be the current time, and let τ_n denote the time step at time t_n. For now, we assume a uniform time partition, so we just use τ. The space approximation at time t_n is given by:

$$u_h^n := \sum_{i\in\mathcal{A}_h} U_i^n \phi_i \in P_k^g(\mathcal{T}_h). \tag{1.40}$$

For the space discretization, we will use forward Euler. The question now is: given u_h^n, (that is, a given set of node values $\{U_i^n\}_{i\in\mathcal{A}_h}$) how do we compute u_h^{n+1} (that is, a new set of node values $\{U_i^{n+1}\}_{i\in\mathcal{A}_h}$)? To do so, we begin by approximating f:

$$f(u_h^n) \approx \sum_{j\in\mathcal{A}_h} f(U_j^n)\phi_j.$$

(This is exact if f is linear—the ϕ_j is scalar valued, and thus can be commuted with f.) Let us define $\mathbf{c} := \int_D \nabla\phi_j \phi_i \mathrm{d}x$. (Observe that $\mathbf{c}$ is vector valued since $\nabla\phi_j$ is vector valued.) Then we have the following:

$$\int_D \nabla\cdot(f(u_h^n))\phi_i \mathrm{d}x \approx \int_D \nabla\cdot\left(\sum_{j\in\mathcal{A}_h} f(U_j^n)\phi_j\right)\phi_i \mathrm{d}x = \sum_{j\in\mathcal{A}_h} f(U_j^n)\cdot\int_D \nabla\phi_j\phi_i\mathrm{d}x = \sum_{j\in\mathcal{A}_j} f(U_j^n)\cdot\mathbf{c}_{ij}.$$

Let us make a few observations about $\mathbf{c}_{ij}$:

- If $i \notin I(j)$ (or $j \notin I(i)$), then $\mathbf{c}_{ij} = \int_D \phi_i \nabla\phi_j \mathrm{d}x = 0$;
- $\mathbf{c}_{ii} = \int_D \phi_i\nabla\phi_i\mathrm{d}x = \int_{\partial D}\phi_i^2\mathbf{n}\mathrm{d}x - \int_D \phi_i\nabla\phi_i\mathrm{d}x = 0 - \int_D \phi_i\nabla\phi_i\mathrm{d}x$. Thus $\mathbf{c}_{ii} = 0$. (Note that are strongly enforcing the boundary conditions, so all our of references functions are 0 on the boundary.) ;
- $\mathbf{c}_{ij} = \int_D \phi_i\nabla\phi_j\mathrm{d}x = \int_{\partial D}\phi_i\phi_j\mathbf{n}\mathrm{d}x - \int_D \phi_j\nabla\phi_i\mathrm{d}x = -\mathbf{c}_{ji}$;
- $\sum_{j\in I(i)}\mathbf{c}_{ij} = \sum_{j\in I(i)}\int_D \phi_i\nabla\phi_j\mathrm{d}x = \int_D \phi_i\nabla\left(\sum_{j\in I(i)}\phi_j\right)\mathrm{d}x = \int_D \phi_i\nabla(1)\mathrm{d}x = 0$. (The sum over $I(i)$ is the same as the sum over $\mathcal{A}_h$ by the first item.)

Now we give the scheme for the update: given $u_h^n \in P_k^g(\mathcal{T}_h)$, we derive $u_h^{n+1} = \sum_{i\in\mathcal{A}_h} U_n^{n+1}\phi_i$ by:

$$m_i \underbrace{\frac{U_i^{n+1}-U_i^n}{\tau}}_{\text{Time discretization}} + \sum_{j\in I(i)}\left(\underbrace{f(U_j)\cdot\mathbf{c}_{ij}}_{\text{Flux in space}} - d_{ij}^n\left(U_j^n - U_i^n\right)\right) = 0, \tag{1.41}$$

where $d_{ij}^n := \left(\max\left(\lambda_{\max}(\mathbf{n}_{ij}, U_i^n, U_j^n)\|\mathbf{c}_{ij}\|, \lambda_{\max}(\mathbf{n}_{ji}, U_j^n, U_i^n\|\right)\mathbf{c}_{ji}\|\right)$ and $\lambda_{\max}$ is any upper bound on the maximum wave speed with data (u_L, u_R) and flux $f\cdot\mathbf{n}_{ij}$. Note:

- $d_{ij}^n = d_{ji}^n$

- d_{ii}^n is irrelevant because the U_i^ns cancel;
- We call d_{ij}^n the *graph viscosity*, which is computed using the $\mathbf{c}_{ij}$, and so is dependent upon ϕ_i.

Remark 1.50. It has been shown that for every nonzero Lipschitz flux, there exists initial data such that the first update violates the maximum principle for every choice of graph viscosity.

Remark 1.51 (Alternative formulation of the update)**.** $\sum_{j\in I(i)} f(U_j^n)\cdot\mathbf{c}_{ij} - d_{ij}^n(U_j^n - U_i^n)$ can be changed to:

$$\sum_{j\in I(i)\setminus\{i\}} (f(U_j^n) - f(U_i^n))\cdot\mathbf{c}_{ij} - d_{ij}^n(U_j^n - U_i^n)$$

since $\mathbf{c}_{ii} = 0$ and

$$\sum_{j\in I(i)\setminus\{i\}} (f(U_j^n) - f(U_i^n))\cdot\mathbf{c}_{ij} = \sum_{j\in I(i)\setminus\{i\}} f(U_j^n)\cdot\mathbf{c}_{ij} - f(U_i^n)\underbrace{\sum_{j\in I(i)}\mathbf{c}_{ij}}_{0} = \sum_{j\in I(i)\setminus\{i\}} f(U_j^n)\cdot\mathbf{c}_{ij}.$$

Thus the scheme is equivalent to:

$$m_i\frac{U_i^{n+1}-U_i^n}{\tau} + \sum_{j\in I(i)\setminus\{i\}} \big(f(U_j^n) - f(U_i^n)\big)\cdot\mathbf{c}_{ij} - d_{ij}^n(U_j^n - U_i^n) = 0. \tag{1.42}$$

Remark 1.52 (Conservation)**.** We calculate:

$$\sum_{i\in\mathcal{A}_h} m_iU_i^n = \sum_{i\in\mathcal{A}_h}\int_D \phi_iU_i^n\mathrm{d}x = \int_D\sum_{i\in\mathcal{A}_h} U_i^n\phi_i\mathrm{d}x = \int_D u_h^n\mathrm{d}x.$$

$$\begin{aligned}\sum_{i\in\mathcal{A}_h} d_{ij}^n(U_j^n - U_i^n) &= \sum_{i\in\mathcal{A}_h}\sum_{j\in I(i)} d_{ij}^nU_j^n - \sum_{i\in\mathcal{A}_h}\sum_{j\in I(i)} d_{ij}^nU_i^n \\ &= \sum_{i\in\mathcal{A}_h}\sum_{j\in I(i)} d_{ij}^nU_j^n - \sum_{j\in\mathcal{A}_h}\sum_{i\in I(j)} d_{ij}^nU_j^n = 0 \text{ since } i\in I(j) \text{ iff } j\in I(i) \text{ and } d_{ij}^n = d_{ji}^n \text{ (work example if confus}\end{aligned}$$

Now we show that the scheme is conservative:

$$\begin{aligned} m_i\frac{U_i^{n+1}-U_i^n}{\tau} &= -\sum_{j\in I(i)}(f(U_j)\cdot\mathbf{c}_{ij} - d_{ij}^n(U_j^n - U_i^n)) \implies \\ m_iU_i^{n+1} &= m_iU_i^n - \tau\sum_{j\in I(i)}(f(U_j^n)\cdot\mathbf{c}_{ij} - d_{ij}(U_j^n - U_i^n)) \implies \\ \sum_{i\in\mathcal{A}_h} m_iU_i^{n+1} &= \sum_{i\in\mathcal{A}} m_iU_i^n - \tau\sum_{i\in\mathcal{A}_j}\sum_{j\in I(i)}(f(U_j)\cdot\mathbf{c}_{ij} - d_{ij}^n(U_j^n - U_i^n)) \\ &= \sum_{i\in\mathcal{A}_j} m_iU_i^n - \tau\sum_{i\in\mathcal{A}_h}\sum_{j\in I(i)} f(U_j^n)\cdot\mathbf{c}_{ij} \implies \\ \int_D u_h^{n+1}\mathrm{d}x &= \int_D u_h^n\mathrm{d}x - \tau\sum_{i\in\mathcal{A}_h}\sum_{j\in I(i)} f(U_j^n)\cdot\mathbf{c}_{ij}. \end{aligned}$$

Calculation shows $\phi_i\nabla\cdot(f(U_j^n)\phi_j) = \phi_i(f(U_j)\cdot\nabla\phi_j)$. Thus:

$$\int_D \phi_i\nabla\cdot(f(U_j^n)\phi_j)\mathrm{d}x = \int_D f(U_j)\phi_i\cdot\nabla\phi_j\mathrm{d}x = f(U_j^n)\cdot\mathbf{c}_{ij}.$$

Therefore,

$$\int_D u_h^{n+1}\mathrm{d}x = \int_D u_h^n \mathrm{d}x - \tau \sum_{i\in\mathcal{A}_h}\sum_{j\in I(i)} \int_D \phi_i f(U_j^n)\cdot\nabla\phi_j \mathrm{d}x.$$

Now,

$$\begin{aligned}\sum_{i\in\mathcal{A}_h}\sum_{j\in I(i)}\int_D \phi_i f(U_j^n)\cdot\nabla\phi_j\mathrm{d}x &= \sum_{j\in I(i)}\int_D f(U_j^n)\cdot\nabla\phi_j\mathrm{d}x\\ &= \sum_{j\in I(i)}\int_D \nabla\cdot(f(U_j^n)\phi_j)\mathrm{d}x\\ &= \int_{\partial D}\sum_{j\in I(i)} f(U_j^n)\phi_j\cdot\mathbf{n}\mathrm{d}x.\end{aligned}$$

By assumption, $U_j^n = U_j^0$ for all i on the boundary. Further assume that $f(U_j^0)\cdot\mathbf{n}_j = 0$, where $n_j = m_j^{-1}\int_{\partial D}\mathbf{n}\phi_j\mathrm{d}x$. Thus, $\int_{\partial D}\sum_{j\in I(i)} f(U_j^n)\phi_j\cdot\mathbf{n}\mathrm{d}x = 0$.
Thus:

$$\int_D u_h^{n+1}\mathrm{d}x = \int_D u_h^n\mathrm{d}x = \cdots = \int_D u_h^0\mathrm{d}x. \tag{1.43}$$

Example 1.53. Let $D=(-1,1)$, $f(x) := f(v)e_x$, $\mathfrak{T}_h$= mesh with cells $\{[x_i, x_{i+1}]\}_{i\in\{1,I-1\}}$ with $x_1 = -1$, $x_I = 1$. Then $\mathcal{A}_h = \{1:I\}$, $\mathcal{A}_h^0 = \{2:I-1\}$, $\mathcal{A}_h^\partial = \{1:I\}$. Our finite element space is $P_1^g(\mathfrak{T}_h)$ = continuous piecewise linear. Computation gives: $\mathbf{c}_{i,i-1} = \frac{-1}{2}$, $\mathbf{c}_{i,i+1} = \frac{1}{2}$, and $m_i = \frac{1}{2}(h_{i-1}+h_i)$, where $h_i := x_{i+1} - x_i$. With these considerations,

$$m_i\frac{U_i^{n+1}-U_i^n}{\tau} + \sum_{j\in I(i)}(f(U_j^n)\cdot\mathbf{c}_{ij} - d_{ij}^n(U_j^n - U_i^n) = 0$$

is such that $I(i)\setminus\{i\} = \{i+1, i-1\}$, and so we conclude:

$$\sum_{j\in I(i)}(f(U_j)\cdot\mathbf{c}_{ij} - d_{ij}^n(U_j^n - U_i^n)) = f(U_{i-1})\cdot\mathbf{c}_{i,i-1} - d_{i,i-1}^n(U_{i-1}-U_i^n) + f(U_{i+1}^n)\cdot\mathbf{c}_{i,i+1} - d^n i, i+1(U_{i+1}^n - U_i^n) =$$

$$\begin{aligned}\sum_{j\in I(i)}(f(U_j)\cdot\mathbf{c}_{ij} - d_{ij}^n(U_j^n - U_i^n)) &= f(U_{i-1})\cdot\mathbf{c}_{i,i-1} - d_{i,i-1}^n(U_{i-1}-U_i^n) + f(U_{i+1}^n)\cdot\mathbf{c}_{i,i+1} - d^n i, i+1(U_{i+1}^n - U_i^n)\\ &= f(U_{i-1})(\frac{-1}{2}) - d_{i,i-1}^n(U_{i-1-U_i}) + f(U_{i+1}^n)\frac{1}{2} - d_{i,i+1}^n(U_{i+1}^n - U_i^n).\end{aligned}$$

Thus,

$$m_i\frac{U_i^{n+1}-U_i^n}{\tau} = \frac{f(U_{i-1}^n) - f(U_{i+1}^n)}{2} + d_{i,i-1}^n(U_{i-1}^n - U_i^n) + d_{ij}^n(U_{i+1}^n - U_i^n).$$

For the purposes of this example, suppose $f(u) = \beta u$. Then, $f'(u) = \beta$, and so $\lambda_{\max}(u_L, u_R) = |\beta|$ for any choice of $u_l,\ u_R$. Furthermore:

$$\begin{aligned}d_{i,i-1} &= \max(\lambda_{\max}(\mathbf{n}_{ij}, U_i^n, U_i^n)\|\mathbf{c}_{ij}\|, \lambda_{\max}(\mathbf{n}_{ij}, U_j^n, U_i^n)\|\mathbf{c}_{ij}\|)\\ &= \max(|\beta|\frac{1}{2}, |\beta|\frac{1}{2})\\ &= \frac{|\beta|}{2}\\ &= d_{i,i+1}.\end{aligned}$$

Thus, this problem becomes:

$$\begin{aligned}
m_i\frac{U_i^{n+1}-U_i^n}{\tau} &= \frac{\beta U_{i-1}^n-\beta U_{i+1}^n}{2}+\frac{|\beta|}{2}(U_{i-1}^n-U_i^n)+\frac{|\beta|}{2}(U_{i+1}^n-U_i^n)\\
&= \frac{1}{2}\left[\beta U_{i-1}^n-\beta U_{i+1}^n+|\beta|U_{i+1}6n+|\beta|U_{i+1}^n-|\beta|U_i^n\right]\\
&= \frac{1}{2}\left[(\beta+|\beta|)U_{i-1}^n-(\beta+|\beta|)U_i^n+(|\beta|-\beta)U_{i+1}^n-(|\beta|-\beta)U_i^n\right] \text{ add } \frac{1}{2}(|\beta|U_i^n-|\beta|U_i^n)=0\\
&= \frac{1}{2}(\beta+|\beta|)(U_{i-1}^n-U_i^n)+\frac{1}{2}(|\beta|-\beta)(U_{i+1}^n-U_i^n).
\end{aligned}$$

Maximum Principle

Define: $\overline{u}(t,\mathbf{n},u_L,u_R):=\int_{-1/2}^{1/2}u(\mathbf{n},u_L,u_R)(x,t)\mathrm{d}x$.

Theorem 1.54. *Let $n\in\mathbb{N}$. Assume that the entries of the lumped mass matrix are positive. Assume that τ is small enough so that the following holds:*

$$\min_{i\in\mathcal{A}_h}\left(1+2\tau\frac{d_{ii}^n}{m_i}\right)\geq 0,$$

where $d_{ii}^n=-\sum_{j\in I(i)\backslash\{i\}}d_{ij}^n$. The following local maximum principle is satisfied: For all $i\in\mathcal{A}_h$:

$$U_i^{n+1}\in[U_i^{m,n},U_i^{M,n}];\ U_i^{m,n}=\min_{j\in I(i)}U_j^n;\ U_j^{M,n}:=\max_{j\in I(i)}U_j^n. \tag{1.44}$$

Proof. By assumption, the result holds for all $i\in\mathcal{A}_h^\partial$, so suppose $i\in\mathcal{A}_h^0$. Recall that $\sum_{j\in I(i)}f(U_j^n)\cdot\mathbf{c}_{ij}=0$ since $\sum_{j\in I(i)}\mathbf{c}_{ij}=0$. Thus:

$$m_i\frac{U_i^{n+1}-U_i^n}{\tau}+\sum_{j\in I(i)}(f(U_j^n)-f(U_i^n))\cdot\mathbf{c}_{ij}-d_{ij}^n(U_j^n-U_i^n)=0$$

implies:

$$\begin{aligned}
U_i^{n+1} &= U_i^n-\frac{\tau}{m_i}\left(\sum_{j\in I(i)}(f(U_j^n)-f(U_i^n))\cdot\mathbf{c}_{ij}-d_{ij}^n(U_j^n-U_i^n)\right)\\
&= U_i^n+\sum_{j\in I(i)\backslash\{i\}}\frac{\tau}{m_i}d_{ij}(U_j^n-U_i^n)-\sum_{j\in I(i)\backslash\{i\}}\frac{\tau}{m_i}(f(U_j^n)-f(U_i^n))\cdot\mathbf{c}_{ij}\\
&= U_i^n-\sum_{j\in I(i)\backslash\{i\}}\frac{2\tau d_{ij}^n}{m_i}U_i^n+\sum_{j\in I(i)\backslash\{i\}}\frac{\tau}{m_i}d_{ij}^nU_j^n-\sum_{j\in I(i)\backslash\{i\}}(f(U_j^n)-f(U_i^n))\frac{\tau}{m_i}\mathbf{c}_{ij}\\
&= \left(1-\sum_{j\in I(i)\backslash\{i\}}\frac{2\tau}{m_i}d_{ij}^n)\right)U_i^n+\sum_{j\in I(i)\backslash\{i\}}\frac{2\tau}{m_i}d_{ij}^n\left[\underbrace{\frac{1}{2}\left(U_i^n+U_j^n-(f(U_j^n)-f(U_i^n))\right)\cdot\frac{\mathbf{c}_{ij}}{2d_{ij}^n}}_{:=\overline{U}_{ij}}\right]\\
&= \left(1-\sum_{j\in I(i)\backslash\{i\}}\frac{2\tau}{m_i}d_{ij}^n)\right)U_i^n+\sum_{j\in I(i)\backslash\{i\}}\frac{2\tau}{m_i}d_{ij}\overline{U}_{ij}.
\end{aligned}$$

Thus, for small enough τ, we that U_i^{n+1} is a convex combination of U_i^n and $\overline{U}_{ij}$. Thus, we only to verify that $\overline{U}_{ij}\in[U_i^n,U_j^n]$ or $\overline{U}_{ij}\in[U_j^n,U_i^n]$. Let $\mathbf{n}_{ij}:=\frac{\mathbf{c}_{ij}}{\|\mathbf{c}_{ij}\|}$ and let $t_{ij}:=\frac{\|\mathbf{c}_{ij}\|}{2d_{ij}^n}$be the "fake time". Then:

$$\overline{U}_{ij}^n = \overline{u}(t_{ij}, \mathbf{n}_{ij}, U_i^n, U_j^n)$$

as in the lemma from the scalar Riemann problem average section, provided $t_{ij}\lambda_{\max}(\mathbf{n}_{ij}, u_L, u_R) \leq \frac{1}{2}$. But:

$$\begin{aligned} t_{ij}\lambda_{\max}(\mathbf{n}_{ij}, U_i^n, U_j^n) &= \frac{\|\mathbf{c}_{ij}\|}{2d_{ij}^n}\lambda_{\max}(\mathbf{n}_{ij}, U_i^n, U_j^n) \\ &\leq \frac{\|\mathbf{c}_{ij}\|}{2\lambda_{\max}(\mathbf{n}_{ij}, U_i^n, U_j^n)\|\mathbf{c}_{ij}\|}\lambda_{\max}(\mathbf{n}_{ij}, U_i^n, U_j^n) \\ &= 1/2 \end{aligned}$$

Thus $\overline{U}_{ij}^n = \overline{u}(t_{ij}, \mathbf{n}_{ij}, U_i^n, U_j^n)$. Being as this is the average of the Riemann problem between $U_i^n,\ U_j^n$, we have that the maximum principle holds. □

For the next corollary, we require (our own) lemma:

Lemma 1.55. *Let $\{\phi_j\}_{j\in\mathcal{A}_h}$ be a partition of unity and $\phi_j(x) \geq 0$ for all j. Let $u_h := \sum_{j\in\mathcal{A}_h} U_j\phi_j(x)$. Denote $U_{\max} := \max_j U_j,\ U_{\min} := \min_j U_j$. Then, $U_{\min} \leq u_h(x) \leq U_{\max}$ for all $x \in D$.*

Proof. Since $\sum_{j\in\mathcal{A}_h}\phi_j(x) = 1$, we have:

$$u_h(x) = \sum_{j\in\mathcal{A}_h} U_j\phi_j(x) \leq \sum_{j\in\mathcal{A}_j} U_{\max}\phi_j(x) = U_{\max};$$

$$U_{\min} = U_{\min}\sum_{j\in\mathcal{A}_h}\phi_j(x) \leq \sum_{j\in\mathcal{A}_j} U_j\phi_j(x) = u_h(x).$$

□

Corollary 1.56. Let $N \in \mathbb{N}\setminus\{0\}$. Assume $\hat{\theta}_i(\hat{x}) \geq 0$ for all $\hat{x} \in \hat{K}$ and all $i \in \mathcal{N}$ and that the above CFL condition is satisfied for all $n < N$. Let $U_{\min}^0 := \min_{j\in\mathcal{A}_h} U_j^0$ and $U_{\max}^0 := \max_{j\in\mathcal{A}_h} U_j^0$. Then:

$$u_h^n(x) \in [U_{\min}^0, U_{\max}^0],\ \forall x \in D,\ n \in \{0:N\}.$$

Proof. Consider $u_h^n(x)$. By the lemma, $U_{\min}^n \leq u_h^n \leq U_{\max}^n$. Now, there exists $j_0,\ j_1 \in \mathcal{A}_h$ such that $U_{\min}^n = U_{j_0}^n,\ U_{\max}^n = U_{j-1}^n$. By the theorem, we have $U_{j_0}^n \in [U_{j_0}^{m,n-1}, U_{j_1}^{M,n-1}]$, $U_{j_1}^n \in [U_{j_1}^{m,n}, U_{j-1}^n M, n]$. But, $U_{\min}^{n-1} \leq U_{j_0}^{m,n-1}, U_{j-1}^{M,n-1} \leq U_{\max}^{n-1}$. Thus, $[U_{\min}^n, U_{\max}^n] \subset [U_{\min}^{n-1}, U_{\max}^{n-1}]$. We may repeat this argument for $n-1, n-2, \cdots, 1$, and thus conclude $u_h^n(x) \in [U_{\min}^n, U_{\max}^n] \subset \cdots \subset [U_{\min}^0, U_{\max}^0]$. □

Entropy Inequalities

Theorem 1.57. *Let $n \in \mathbb{N}$. Assume that the CFL condition holds. Let $(\eta, \mathbf{q})$ be an entropy pair. Then the following inequality holds for all $i \in \mathcal{A}_h$:*

$$\frac{m_i}{\tau}\left(\eta(U_i^{n+1} - \eta((U_i^n)\right) + \int_D \nabla\cdot\left(\sum_{j\in I(i)}\mathbf{q}(U_j^n)\phi_j\right)\phi_i \mathrm{d}x - \sum_{j\in I(i)} d_{ij}^n\left(\eta(U_j^n) - \eta(U_i^n)\right) \leq 0.$$

Proof. Recall that the computation the CFL condition implies that U_i^{n+1} is a convex combination. Thus:

$$\eta(U_i^{n+1}) \leq \left(1 - \sum_{j\in I(i)\setminus\{i\}}\frac{2\tau d_{ij}^n}{m_i}\right)\eta(U_i^n) + \sum_{j\in I(i)\setminus\{i\}}\frac{2\tau d_{ij}^n}{m_i}\eta(\overline{U}_{ij}^n).$$

By assumption, f is Lipschitz, and so we may apply the Riemann average for the entropy inequalities. Thus:

$$\eta(\overline{U}_{ij}^n) \leq \frac{1}{2}(\eta(U_i^n) - \eta(U_j^n)) - t_{ij}(\mathbf{q}(U_j^n)\cdot \mathbf{n}_{ij} - \mathbf{q}(U_i^n)\cdot \mathbf{n}_{ij}).$$

Rearranging the earlier terms and using this inequality gives:

$$\frac{m_i}{\tau}((\eta(U_i^{n+1}) - \eta(U_i 6n)) \leq \sum_{j\in I(i)\setminus\{i\}} 2d_{ij}^n(\eta(\overline{U}_{ij}^n) - \eta(U_n^n)) \leq \sum_{j\in I(i)\setminus\{i\}} \left[d_{ij}^n(\eta(U_j^n) - \eta(U_i^n)) - \|\mathbf{c}_{ij}\|(\mathbf{q}(U_j^n) - \mathbf{q}(U_i^n)\cdot \mathbf{c}_{ij}\right].$$

Thus,

$$\begin{aligned}
\frac{m_i}{\tau}\left(\eta(U_i^{n+1} - \eta(U_i^n)\right) &\leq \sum_{j\in I(i)\setminus\{i\}} 2d_{ij}^n(\eta(\overline{U}_{ij}^n) - \eta(U_i^n)) \\
&\leq \sum_{j\in I(i)\setminus\{i\}} 2d_{ij}^n \left(\frac{1}{2}(\eta(U_i^n) + \eta(U_i^n)) - \underbrace{t_{ij}}_{t_{ij}=\frac{\|c_{ij}\|}{2d_{ij}^n}} (q(U_j^n)\cdot n_{ij} - q(U_i^n)\cdot n_{ij})\right) \\
&= \sum_{j\in I(i)\setminus\{i\}} \left(d_{ij}(\eta(U_i^n) + \eta(U_i^n)) - \|c_{ij}\|(q(U_j^n)\cdot n_{ij} - q(U_i^n)\cdot n_{ij})\right).
\end{aligned}$$

The result follows from the definition of $\mathbf{n}_{ij}$, $\mathbf{c}_{ij}$ and that the sum can be extended to the whole space. □

1.4.2 Hyperbolic Systems First Order Approximation

Let us first develop our setting:

- $n \in \mathbb{R}^d$, normal vector
- (u_L, u_R), data with $\lambda_{\max}(n, u_L, u_R)$ a bound on wave speed
- Our FEM space:
$$\mathbf{P}_k^g(\mathcal{T}) := (P_k^g(\mathcal{T}_h))^m$$
- $\{\phi_i\}_{i\in\mathcal{A}_j}$ is a basis for $P_k^g(\mathcal{T}_h)$.
- $\{e_k\}_{k\in\{1:m\}}$, canonical basis for $\mathbb{R}^m$
- We build a new basis: $\{e_k\phi_i\}_{k\in\{1:m\},i\in\mathcal{A}_h}$.

The Scheme:

Let $u_h^0(x) := \sum_{i\in\mathcal{A}_h} U_i^0\phi_i(x)$ be a "reasonable" approximation of $u_0(x)$. As in the scalar case, we suppose we have $u_h^n(x) := \sum_{i\in\mathcal{A}_h} U_i^n\phi_i(x)$. We compute the update u_h^{n+1} via:

$$m_i\frac{U_i^{n+1} - U_i^N}{\tau} + \sum_{j\in I(i)} \left(f(U_j^n)\cdot \mathbf{c}_{ij} - d_{ih}^n(U_j^n - U_i^n)\right) = 0. \tag{1.45}$$

As before, $\mathbf{c}_{ij} = \int_D \phi_i\nabla\phi_j \mathrm{d}x$; $d_{ij} := \max(\lambda_{\max}(n_{ij}, U_i, U_j)\|\mathbf{c}_{ij}\|, \lambda_{\max}(n_{ij}, U_j, U_i)\|\mathbf{c}_{ij}\|)$, with $n_{ij} := \mathbf{c}_{ij}/\|c_{ij}\|$.

For now, we are concerned with whether or not the scheme is invariant (the system analog for maximum principle). Throughout, we assume that $B \subset A$ is convex.

Theorem 1.58. *Let $n \in \mathbb{N}$. Assume that $m_i > 0$ for all $i \in \mathcal{A}_h$.*

1. *Under the CFL condition $\min_{i\in\mathcal{A}_h}\left(1 + 2\tau\frac{d_{ij}^n}{m_i}\right) \geq 0$, we have:*

$$[\{U_i^n\}_{i\in\mathcal{A}_h} \subset \overline{B}] \implies [\{U_i^{n+1}\}_{i\in\mathcal{A}_h} \subset \overline{B}];$$

2. Under the condition $\min_{i\in\mathcal{A}_h}\left(1+2\tau\frac{d_{ij}^n}{m_i}\right)>0$ *we have:*

$$[\{U_i^n\}_{i\in\mathcal{A}}\subset int(B)]\implies[\{U_i^{n+1}\}_{i\in\mathcal{A}_h}\subset int(B)].$$

Proof. Similarly as before, we can rewrite the scheme as:

$$U_i^{n+1}=\left(1-\sum_{j\in I(i)}2\tau\frac{d_{ij}^n}{m_i}\right)U_i^n+\sum_{j\in I(i)}\overline{U}_{ij}^n.$$

Now, by lemma [2] (Riemann average for hyperbolic systems), with these CFL conditions, we obtain that $\overline{U}_{ij}^n\in\overline{B}$ for all j. Thus $U_i^{n+1}\in\overline{B}$. Finally, since this holds for all i, $\{U_i^{n+1}\}_{i\in\mathcal{A}_h}\subset\overline{B}$. For the second statement, we repeat the proof, but with $int(B)$ instead of $\overline{B}$. □

As in the scalar, we would like the full approximating function to be in the convex set too.

Corollary 1.59. Let $N\in\mathbb{N}\setminus\{0\}$. Assume that the reference shape functions are such that $\theta_i\geq 0$ for all $i\in\mathcal{N}$.

1. Assume that $\min_{i\in\mathcal{A}_h}\left(1+2\tau\frac{d_{ij}^n}{m_i}\right)\geq 0$ for all $n<N$ and that $\{U_i^0\}_{i\in\mathcal{A}_h}\subset\overline{B}$. Then, $\{U_i^n\}_{i\in\mathcal{A}_h}\subset\overline{B}$ and $u_h^n(x)\in\overline{B}$ for all $n\in\{0:N\}$ and $x\in D$.
2. Assume that $\min_{i\in\mathcal{A}_h}\left(1+2\tau\frac{d_{ij}^n}{m_i}\right)>0$ for all $n<N$ and that $\{U_i^0\}_{i\in\mathcal{A}_h}\subset int(B)$. Then $\{U_i^n\}_{i\in\mathcal{A}_h}\subset int(B)$ and $u_h^n(x)\in int(B)$ for all $x\in D$, $n\in\{0:N-1\}$.

Proof. Let $n<N$. Recall that $u_h^n(x)=\sum_{i\in\mathcal{A}_h}U_i^n\phi_i(x)$. Because $\phi_i(x)\geq 0$ for all $i\in\mathcal{A}_h$, we conclude that u_h^n is a convex combination of $\{U_i^n\}_{i\in\mathcal{A}}$. Be repeated applications of the above theorem, we conclude that $\{U_i^n\}_{i\in\mathcal{A}_h}\subset\overline{B}$. Thus, since B is convex, $u_h^n(x)\in\overline{B}$ for all $x\in D$. The second statement is a similar proof but with $int(B)$ replacing $\overline{B}$. □

1.5 Chapter 82: Higher order approximation

With the first order approximation complete, we want to increase our accuracy in both space and time. We first address time and then address space (but space only for scalar valued equations).

1.5.1 Higher order in time

For time accuracy, we use a method called *contractive* or *strong stability preserving* (SSP). Specifically, we use SSP and explicit Runge-Kutta (ERK), i.e., SSPRK methods. For time, the basic idea is to create convex combinations of forward Euler steps such that each step has the invariant domain property. Simply put, each step (or update) is decomposed into sub-steps (or sub-updates) with the final update being a convex combination of the intermediate steps. Let us briefly recall the algorithm (recall that we are still working in the general case):

$$m_i\frac{\mathbf{U}_i^{n+1}-\mathbf{U}_i^n}{\tau}+\sum_{j\in I(i)}\left((\mathbf{f}(\mathbf{U}_j^n)\cdot\mathbf{c}_{ij}-d_{ij}^n(\mathbf{U}_j^n-\mathbf{U}_i^n)\right)=0.$$

This can be rewritten as:

$$\mathbf{U}_i^{n+1}=\mathbf{U}_i^n-\frac{\tau}{m_i}\left(\sum_{j\in I(i)}\left(\mathbf{f}(\mathbf{U}_j^n)-d_{ij}^n(\mathbf{U}_j^n-\mathbf{U}_i^n)\right)\right.$$

For the sake of good mathematics, let us generalize a bit. Let us define the operator $L:\mathbb{R}^{m\times I}\mapsto\mathbb{R}^{m\times I}$ such that for all $i\in\mathcal{A}_i=\{1:I\}$:

$L(\mathbf{U})_i \in \mathbb{R}^m$ is given by $L(U)_i = \dfrac{1}{m_i} \sum_{j \in I(i)} \left(f(\mathbf{U}_j) \cdot \mathbf{c}_{ij} - d_{ij}(\mathbf{U}_j^n - \mathbf{U}_i^n) \right).$

Thus, the one-step algorithm is given by:

$$\mathbf{U}^{n+1} = \mathbf{U}^n - \tau L(\mathbf{U}^n). \tag{1.46}$$

To be careful with notation, we have:

$$\mathbf{U}^{n+1} = [\mathbf{U}_1^n, \dots, \mathbf{U}_I^n].$$

That is, a vector of vectors and L take these objects from one space to itself.

From the previous chapter, under the CFL condition $2\tau \max_{i \in \mathcal{A}_h} \frac{|d_{ij}^n|}{m_i} < 1$, we have:

$$[\{\mathbf{U}_i^n\}_{i \in \mathcal{A}} \subset int(B)] \implies \left[\{\mathbf{U}\}_{i \in \mathcal{A}}^{n+1} \subset int(B)\right].$$

Thus, $int(B)^I$ is invariant under the operator $I - \tau L$. Recall that the same holds for $\overline{B}$ when the CFL condition is relaxed to be $2\tau \frac{|d_{ij}^n|}{m_i} \leq 1$.

This is the property we want to preserve, but with higher accuracy in time. To continue the generality, let E be a finite dimensional vector space, $A \subset E$ and time horizon $T > 0$. Consider an operator $L : [0, T] \times A \to E$. We want to approximate the time-evolution problem: $\partial_t u + L(t, u) = 0$. (For this to make sense, we assume L is continuous in time and Lipschitz in u.) We also make a critical assumption: that there exists a $\tau^* > 0$, $B \subset A$, B convex, such that for all $t \in [0, T]$, $s \in [0, \tau^*]$:

$$[v \in B] \implies [v + sL(t, v) \in B].$$

For the sake of clarity, let give some notation for the SSPRK method.

1. α_{ik} β_{ij}, $i \in \{1 : s\}$, $k \in \{0 : i - 1\}$, i.e., $1 \leq k + 1 \leq s$.
2. $\{c_i\}_{i \in \{1:s\}}$ from the usual SSPRK method, but shift to $\{\gamma_k\}_{k \in \{0:s-1\}}$.

Now we introduce the SSPRK method we will invoke: For all $n \in \mathcal{N}$.

1. Given $u^n \in A$, set $w^0 = u^n$. Compute $\{w^i\}_{i \in \{1:s\}}$ by: (this is the "sub-update")

$$w^i = \sum_{k \in \{0:i-1\}} \alpha_{ik} w^k + \beta_{ik} \tau L(t_n + \gamma_k \tau, w^k) \text{ for all } i \in \{1 : s\}.$$

2. The update u^{n+1} is given by $u^{n+1} = w^s$. (So we have s sub-updates.)
3. We require $\sum_{k \in \{0:i-1\}} \alpha_{iK} = 1$, $\alpha_{ik}, \beta_{ik} \geq 0$ and $\alpha_{ik} = 0 \implies \beta_{ik} = 0$.
4. With these considerations, we can rewrite the sub-update as (with $K_i = \{k \in \{1 : i - 1\} : \ \alpha_{ik} \neq 0\}$)

$$w^i = \sum_{k \in K_i} \alpha_{ik} \left(w^k + \frac{\beta_{ij}}{\alpha_{ik}} \tau L(t_n + \gamma_k \tau, w^k) \right). \tag{1.47}$$

Thus, observe that each sub-update is a convex combination of the previous (forward-Euler made) sub-updates. Lastly, define:

$$c_{os} = \inf_{i \in \{1:s\}} \inf_{k \in K_i} \alpha_{ik} \beta_{ik}^{-1}; \quad \beta_{ik} = 0 \implies \alpha_{ik} \beta_{ij}^{-1} = \infty.$$

We can now give a CFL condition that guarantees this higher order algorithm satisfies the invariant domain properties.

Theorem 1.60. *Let the SSPRK method be defined as above with coefficients that satisfy the listed conditions. Let $B \subset E$ be a convex set and assume that there is a τ^* such that invariance condition above holds. Let c_{os} be defined as above. Then the following holds true for all $\tau \leq c_{os}\tau^*$:*

$$[u^n \in B] \implies \left[u^{n+1} \in B\right].$$

Proof. Suppose $u^n \in B$. We proceed via induction on $i \in \{0:s\}$ that $w^i \in B$. Clearly $w^0 = u^n \in B$ implies that $w^0 \in B$. Now let $i \in \{0:s\}$ and suppose that $w^k \in B$ for all $k \in \{0:i-1\}$. Let

$$z^{i,k} := w^k + \alpha_{ik}^{-1}\beta_{ik}\tau L(t_n + \gamma_k\tau, w^k) \quad k \in K_i.$$

Then, the sub-update can we written as:

$$w^i = \sum_{k\in K_i} \alpha_{ik} z^{i,k}.$$

As noted before, we thus have that w^i is a convex combination, in particular of the z^{ik}s. Thus, we if can ascertain that each $z^{ik} \in B$, we deduce that $w^i \in B$. To this end, by assumption, $\tau \leq c_{os}\tau^*$. Thus, $\tau \leq \tau^*\beta_{ik}^{-1}\alpha_{ik}$ for all $i \in \{1:s\}$ and all $k \in K_i$. In particular, the invariance condition for $v + sL(t,v)$ holds. Thus, since $w^k \in B$ (by assumption), we have that $z^{ik} \in B$. Therefore $w^i \in B$. This holds for all $i \in \{0:s\}$, in particular for $i = s$, and so $u^{n+1} = w^s \in B$. □

1.5.2 Higher order in space for scalar systems

The main idea is to reduce the fake viscosity in areas for the where the value is far away from the local maxima. That is, we reduce viscosity in regions where the maximum principle is not in danger of being violated. In areas where the principle is in danger, ie, the value is close to the maximum, we keep the first order algorithm from previous to keep the maximum principle.

Recall the algorithm:

$$m_i\frac{U_i^{n+1} - U_i^n}{\tau} + \sum_{j\in I(i)} \left[f(U_j^n)\cdot c_{ij} - d_{ij}^n(U_j^n - U_i^n)\right] = 0; \text{ or}$$

$$U_i^{n+1} = U_i^n + \frac{\tau}{m_i}\left[\sum_{j\in I(i)} \left(f(U_j^n)\cdot c_{ij} - d_{ij}^n(U_j^n - U_i^n)\right)\right].$$

where $d_{ij}^n = \max(\lambda_{\max}(n_{ij}, U_i^n, U_j^n)\|c_{ij}\|, \lambda_{\max}(n_{ij}, U_j^n, U_i^n))$, $n_{ij} = \frac{c_{ij}}{\|c_{ij}\|}$, and $c_{ij} = \int_D \phi_i \nabla\phi_j \mathrm{d}x$.

This is the first order algorithm. Because of this fact, let us denote for the higher order algorithm:

$$d_{ij}^n := d_{ij}^{L,n}.$$

That is, for the higher order algorithm, the "L" denotes the fact that we are suing the lower order form.

Now, denote for all n a collection of weights:

$$\psi_i^n \in [0,1], \quad i \in \mathcal{A}_h.$$

With these weights, we can define the higher order viscosity terms:

$$d_{ij}^n := d_{ij}^{L,n}\max(\psi_i^n, \psi_j^n), \quad j \in I(i)\setminus\{i\}.$$

We keep the notation:

$$d_{ii}^n := -\sum_{j\in I(i)\setminus\{i\}} d_{ij}^n. \text{ (but with the new } d_{ij}^n \text{ above.}$$

The big question now is how to choose the $\{\psi_i^n\}_{i\in\mathcal{A}_h}$ carefully so that the maximum principle is satisfied. That, we want:

$$U_i^{n+1} \in [U_i^{m,n}, U_i^{m,n}], \quad U_i^{m,n} = \min_{j \in I(i)} U_j^n, \quad U_i^{m,n} = \min_{j \in I(i)} U_j^n.$$

We define a local CFL number based on the low- order viscosity:

$$\gamma_i^n := \frac{2\tau |d_{ij}^{L,n}}{m_i}, \quad d_{ii}^{L,n} := -\sum_{j \in I(i) \setminus \{i\}} d_{ij}^{L,n}.$$

Quick observation: if all the weights are equal to 1, ie, $\psi_i^n = 1$ for all i and n, then $d_{ij}^n = d_{ij}^{L,n}$, and we resort to the low order algorithm. Then, under the CFL condition:

$$\max_{i \in \mathcal{A}_h} \gamma_i^n = \max_{i \in \mathcal{A}_h^0} \frac{2\tau |d_{ii}}{m_i} \leq 1,$$

we have:

$$\min_{i \in \mathcal{A}_h^0} \left(1 + \frac{2\tau d_{ij}}{m_i}\right) \geq 0.$$

Thus the CFL condition holds for the maximum principle.

For the case when $\psi_i^n \neq 0$, we use the following:

Define $\theta_i^n \in [0,1]$ such that:

$$\theta_i^n := \frac{U_i^n - U_i^{m,n}}{U_i^{M,n} - U_i^{m,n}} \text{ if } U_i^{M,n} - U_{[}^{m,n}] \neq 0, \quad \frac{1}{2} \text{ otherwise.}$$

Quickly observe that this definition implies

$$U_i^n = \theta_i^n U_i^{M,n} + (1 - \theta_i^n) U_i^{m,n}.$$

Let us define the following sets:

$$I(i^+) := \{j \in I(i) : \ U_i^n < U_j^n\}; \quad I(i^-) := \{j \in I(i) : \ U_i^n > U_j^n\}.$$

Further define:

$$\gamma_i^{+,n} := \frac{2\tau}{m_i} \sum_{j \in I(i^+)} d_{ij}^{L,n}; \quad \gamma_i -, n := \frac{2\tau}{m_i} \sum_{j \in I(i^-)} d_{ij}^{L,n}.$$

If $I(i^\pm) = \emptyset$, set $\gamma_i^{\pm,n} = 0$.

The first result is of serious importance, and it gives a tool for not just achieving the maximum principle, but also a stronger estimate on the update.

Lemma 1.61. *Let $n \geq 0$ and $i \in \mathcal{A}_h^0$. Assume that $\gamma_i^n < 1$ and $U_i^{M,n} - U_i^{m,n} \neq 0$. Define:*

$$\delta_i^{M,n} := (1 - \theta_i^n)(1 - \gamma_i^n) - \theta_i^n (1 - \psi_i^n) \frac{1}{2} \gamma_i^{-,n}$$
$$\delta_i^{m,n} := \theta_i^n (1 - \gamma_i^n) - (1 - \theta_i^n)(1 - \psi_i^n) \frac{1}{2} \gamma_i^{+,n}.$$

Let U_i^{n+1} be defined as in the algorithm. Then:

$$U_i^{n+1} \in \left[U_i m, n + (U_i^{M,n} - U_i^n)\delta_i^{m,n}, U_i^{M,n} - (U_i^{M,n} - U_i^n)\delta_i^{M,n}\right].$$

Smoothness based graph viscosity

The object of this section is to develop a way to pick the weight $\{\psi_j^n\}$ such that the maximum principle holds. The moral of the story here is to find good conditions such that the weights are small (so we have smaller viscosity) but we preserve the maximum principle.

Suppose $U_i^{M,n} \neq U_i^{m,n}$. Introduce the following object:

$$\alpha_i^j := \frac{\left|\sum_{j\in I(i)} \beta_{ij}(U_j^n - U_i^n)\right|}{\sum_{j\in I(i)} \beta_{ij}\left|U_j^n - U_i^n\right|}; \quad \beta_{ij} \geq 0, \quad \beta_{ij} \neq 0 \text{ for at least one } ij.$$

The idea is to get the weights (ψ_i^n) from a function

$$\psi \in Lip([0,1],[0,1]) \quad w/ \quad \psi(1) = 1, \quad \psi_i^n := \psi(\alpha_i^n).$$

If $U_i^n =$ local extremum (so min or max), then have that $\alpha_j^n = 1$ (check if in doubt). Thus, when U_i^n is a local max or min, $\alpha_i^n = 1$ so $\psi_i^n = 1$, and as previously discussed, this resorts to the low order algorithm.

In general, if $\psi(\alpha) := 1$, then we resort to the low order method, which satisfies the maximum principle. Let us define:

$$\beta_i^m := \min_{j\in I(i)} \beta_{ij}; \quad \beta_i^M := \max_{j\in I(i)} \beta_{ij}.$$

Furthermore, suppose we have a $\beta_\#$ such that for all h:

$$0 < \beta_{ij} \text{ for all } i \in \mathcal{A}_h,\, j \in I(i)\ ; \quad \max_{i\in\mathcal{A}_h} \frac{\beta_i^M}{\beta_i^m} \leq \beta_\#.$$

Finally set:

$$c_\# := \beta_\# \max_{i\in\mathcal{A}_h} card(I(i)).$$

We these assumptions, we can prove a maximum principle.

Theorem 1.62. *Let $\psi \in Lip([0,1],[0,1])$ be such that $\psi(1) = 1$. Let K_ψ denote the Lipschitz constant for ψ. Set $\psi_i^n := \psi(\alpha_i^n)$ for all $i \in \mathcal{A}_h$ and $n \geq 0$ with α_i^n, β_{ij} defined as above. Then the algorithm with the new d_{ij}^n satisfies the local maximum principle under the CFL condition $\max_{i\in\mathcal{A}_h} \gamma_i^n \leq 1/(1 + k_\psi c_\#)$.*

Proof. First, observe that if $U_i^{M,n} = U_i^{m,n}$, then $U_i^n = U_j^n$ for all $j \in I(i)$. If we recall the scheme:

$$m_i \frac{U_i^{n+1} - U_i^n}{\tau} + \sum_{j\in I(i)\setminus\{i\}} \left[f(U_j)\cdot c_{ij} - d_{ij}^n(U_j^n - U_i^n)\right] = 0,$$

we see that the d_{ij}^n terms vanish since $U_j^n = U_i^n$ and that

$$\sum_{j\in I(i)\setminus\{i\}} f(U_j)\cdot c_{ij} = f(U_i) \sum_{j\in I(i)\setminus\{i\}} c_{ij} = f(U_i^n)0 = 0.$$

Thus, $U_i^{n+1} = U_i^n \in [U_i^{m,n}, U_i^{M,n}]$ and the maximum principle is satisfied.

Now, if $\theta_i^n \in \{0,1\}$, then $U_i^n = U_i^{m,n}$ or $U_i^{M,n}$ respectively. As previously noted, this implies that the algorithm resorts to the first order version, and thus satisfies the maximum principle. Finally, suppose that $\theta_i^n \in (0,1)$. Let us define the objects:

$$s_i^+ = \sum_{j\in I(i^+)} \beta_{ij}\left|U_j^n - U_i^n\right|; \quad s_i^- = \sum_{j\in I(i^-)} \beta_{ij}\left|U_j^n - U_i^n\right|.$$

Note that $i \notin I(i^+)$, $i \notin I(i^-)$. With these objects, we observe:

$$s_i^+ = \sum_{j\in I(i^+)} \beta_{ij}(U_j^n n - U_i^n), \quad s_i^- = \sum_{j\in I(i^-)} -\beta_{ij}(U_j^n - U_i^n),$$

and that:

$$s_i^+ - s_i^- = \sum_{j \in I(i) \setminus \{i\}} \beta_{ij}(U_j^n - U_i^n); \quad s_i^+ + s_i^- = \sum_{j \in I(i) \setminus \{i\}} \beta_{ij} \left| U_j^n - U_i^n \right|.$$

Thus, we compute:

$$\begin{aligned} 1 - \alpha_i^n = 1 - \frac{\left| s_i^+ - s_i^- \right|}{s_i^+ - s_i^-} &= \frac{s_i^+ - s_i^-}{s_i^+ - s_i^-} - \frac{\left| s_i^+ - s_i^- \right|}{s_i^+ - s_i^-} \\ &\leq \frac{s_i^+ - s_i^-}{s_i^+ - s_i^-} + \frac{s_i^+ - s_i^-}{s_i^+ - s_i^-} \\ &= 2\frac{s_i^+}{s_i^+ - s_i^-} \\ &= 2\frac{\sum_{j \in I(i^+)} \beta_{ij}(U_j^n - U_i^n)}{\sum_{j \in I(i)} \beta_{ij} \left| U_j^m - U_i^n \right|}. \end{aligned}$$

Note the following inequalities:

$$\begin{aligned} \beta_{j \in I(i)} \beta_{ij} \left| U_j^n - U_i^n \right| &\leq \sum_{j \in I(i)} \beta_{ij} \left| U_j^{m,n} - U_i^n \right| \\ \sum_{j \in I(i)} \beta_{ij} \left| U_j^n - U_i^n \right| &\geq \sum_{j \in I(i)} \beta_i^m \left| U_j^n - U_i^n \right| \\ &\geq \beta_i^m (\left| U_i^{M,n} - U_i^n \right| + |U_i^{m,n} - U_i^n|). \end{aligned}$$

Thus:

$$\begin{aligned} 2\frac{\sum_{j \in I(i)} \beta_{ij}(U_j^n - U_i^n)}{\sum_{j \in I(i)} \beta_{ij} \left| U_j^n - U_i^n \right|} &\leq 2\frac{\sum_{j \in I(i^+)} \beta_{ij}(U_j^{M,n} - U_i^n)}{\beta_i^m (\left| U_i^{M,n} - U_i^n \right| + |U_i^{m,n} - U_i^n|)} \\ &\leq 2\frac{U_j^{M,n} - U_i^n}{U_i^{M,n} - U_i^m} \frac{\beta_i^M}{\beta_i^m} card(I(i^+)) \\ &= 2(1 - \theta_i^n)\frac{\beta_i^M}{\beta_i^m} card(I(i^+)) \\ &\leq 2(1 - \theta_i^n) c_\#. \end{aligned}$$

A similar style argument gives that $1 - \alpha_i^n \leq 2\theta_i^n c_\#$. We thus deduce:

$$\begin{aligned} 1 - \psi(\alpha_i^n) = \psi(1) - \psi(\alpha_i^n) &\leq k_\psi(1 - \alpha_i^n \\ &\leq k_\psi c_\# \min(\theta_i^n, 1 - \theta_i^n). \end{aligned}$$

(Readily observe that $\gamma_i^{-,n} \leq \gamma_i^n$ by definition.) Therefore:

$$\begin{aligned} \delta_i^{M,n} &= (1 - \theta_i^n)(1 - \gamma_i^n) - \theta_i^n(1 - \psi(\alpha_i^n))\frac{1}{2}\gamma_i^{-,n} \\ &\geq (1 - \theta_i^n)(1 - \gamma_i^n) - \theta_i^n(1 - \theta_i^n)k_\psi c_\# \gamma_i^n \\ &= (1 - \theta_i^n)\left[(1 - \gamma_i^n) - \theta_i^n k_\psi c_\# \gamma_i^n\right] \\ &= (1 - \theta_i^n)\left[1 - (1 + k_\psi c_\# \theta_i^n)\right]. \end{aligned}$$

The last quantity is non-negative when

$$(1 + k_\psi c_\# \theta_i^n)\gamma_i^n \leq 1, \text{ that is, when } \gamma_i^n \leq \frac{1}{1 + k_\psi c_\#}.$$

Similarly, we deduce:

$$\begin{aligned}
\delta_i^{m,n} &= \theta_i^n(1 - \gamma_i^n) - (1 - \theta_i^n)(1 - \psi(\alpha_i^n))\frac{1}{2}\gamma_i^{+,n} \\
&\geq \theta_i^n(1 - \gamma_i^n) - (1 - \theta_i^n)\theta_i^n k_\psi c_\# \gamma_i^n \\
&= \theta(1 - \gamma_i^n - (1 - \theta_i^n)k_\psi c_\# \gamma_i^n) \\
&= \theta_i^n(1 - (1 + (1 - \theta_i^n)k_\psi c_\#)\gamma_i^n).
\end{aligned}$$

As before, this quantity is non-negative when

$$\gamma_i^n \leq \frac{1}{1 + k_\psi c_\#}.$$

This holds by assumption by the CFL condition. Thus, $\delta_i^{M,n}, \delta_i^{m,n} \geq 0$, and thus by the lemma, we satisfy the maximum principle. □

Greedy Graph Viscosity

The main idea of this approach is to satisfy the gap lemma without invoking any notion of smoothness.

Theorem 1.63. *Let θ_i^n, γ_i^n, $\gamma_i^{\pm,n}$ be defined as before. Define the weights:*

$$\psi_i^n := \max\left(1 - 2(1 - \gamma_i^n)\min\left(\frac{1}{\gamma_i^n}\frac{1 - \theta_i^n}{\theta_i^n}, \frac{1}{\gamma_i^{+,n}}\frac{\theta_i^n}{1 - \theta_i^n}\right), 0\right)$$

with the convention that $\psi_i^n - 1$ if $\theta_i^n \in \{0, 1\}$. Then the (high - order) space algorithm satisfies the maximum principle under the CFL condition $\max_{i \in \mathcal{A}_h} \gamma_i^n \leq 1$.

Proof. AS before, if $U_i^{m,n} = U_i^{M,n}$, then $U_i^{n+1} = U_i^n \in [U_i^{m,n}, U_i^{M,n}]$, which implies the maximum holds under the CFL condition. If $\theta_i^n \in \{0, 1\}$, then $\psi_i^n = 1$. Then, since $\gamma_i^n \leq 1$ by assumption, $d_{ij}^n = d_{ij}^{L,n}$ for all $j \in I(i)$. We thus resort to the low-order algorithm, and since this is a tighter CFL condition, we achieve the maximum principle.

Lastly, assume $\theta_i^n \in (0, 1)$. By assumption, we get that:

$$\begin{aligned}
\psi_i^n &\geq 1 - 2(1 - \gamma_i^n)\min\left(\frac{1}{\gamma_i^n}\frac{1 - \theta_i^n}{\theta_i^n}, \frac{1}{\gamma_i^{+,n}}\frac{\theta_i^n}{1 - \theta_i^n}\right) \\
&\geq 1 - \frac{2(1 - \gamma_i^n)}{\gamma_i^{-,n}}\frac{1 - \theta_i^n}{\theta_i^n}.
\end{aligned}$$

Thus:

$$\begin{aligned}
\delta_i^{M,n} &= (1 - \gamma_i^n)(1 - \theta_i^n) + \theta_i^n(\psi_i^n - 1)\frac{\gamma_i^{-,n}}{2} \\
&\geq (1 - \gamma_i^n)(1 - \theta_i^n) + \theta_i^n\left(1 - \frac{2(1 - \gamma_i^n)}{\gamma_i^{-,n}}\frac{1 - \theta_i^n}{\theta_i^n} - 1\right)\frac{\gamma_i^{-,n}}{2} \\
&= (1 - \gamma_i^n)(1 - \theta_i^n) + \theta_i^n\left(-\frac{2(1 - \gamma_i^n)}{\gamma_u^{-,n}}\frac{1 - \theta_i^n}{\theta_i^n}\right)\frac{\gamma_i^{-,n}}{2} \\
&= (1 - \gamma_i^n)(1 - \theta_i^n) + (-(1 - \theta_i^n)(1 - \theta_i^n)) \\
&= 0.
\end{aligned}$$

A similar argument gives that $\delta_i^{m,n} \geq 0$ as well. Thus, after applying the gap lemma, we conclude that the maximum principle is satisfied. $\square$