Numerical Methods for Hyperbolic Conservation Laws (AM257)

BY CHI-WANG SHU

Semester I 2006, Brown. Send corrections to `kloeckner@dam.brown.edu`. Any mistakes or omissions in these notes are certainly due to my typing.

Table of contents

1 Theory of One-Dimensional Scalar Conservation Laws

$$u_t + f(u)_x = 0, \tag{1}$$

where u is a function of x and t.

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_a^b u(x,t)\mathrm{d}x = f(u(b,t)) - f(u(a,t))$$

is the integral form of (1).

$$\begin{cases} u_t + f(u)_x, \\ u(x,0) = u^0(x). \end{cases} \tag{2}$$

Characteristics: Define a function $x(t)$ by

$$\begin{cases} \frac{\mathrm{d}x(t)}{\mathrm{d}t} = f'(u(x(t),t)), \\ x(0) = x_0. \end{cases}$$

Then

$$\frac{\mathrm{d}u(x(t),t)}{\mathrm{d}t} = u_x x'(t) + u_t = u_x f'(u(x(t),t) + u_t = f(u)_x + u_t = 0.$$

So $u(x(t),t) = u(x(0),0) = u^0(x_0)$.

All that holds under the assumption that we have a smooth solution. Which we don't. :(

Consider *Burgers' Equation*:

$$\begin{cases} u_t + \left(\frac{u^2}{2}\right)_x = 0, \\ u(x,0) = \sin(x). \end{cases} \tag{3}$$

Consider the characteristics at $\pi/2$ and $3\pi/2$. → They intersect and propagate different values, so the above theory breaks down. ⇒ There is no global (in x and t) solution to (3). The concept of "weak solution" helps us out now. Reconsider the integral form:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b u(x,t)\mathrm{d}x = f(u(a,t)) - f(u(b,t)) \tag{4}$$

For C^1 solutions, (1) ⇔ (4). Attempts at defining weak solutions:

- If u satisfies (4) for almost all (a, b) then in u is called a weak solution to (1). (physically meaningful, correct)
- If for any $\varphi \in C_0^1(\mathbb{R}^2)$,

  $$-\int_0^\infty \int_{-\infty}^\infty (u\varphi_t + f(u)\varphi_x)\mathrm{d}x\,\mathrm{d}t - \int_{-\infty}^\infty u^0(x)\varphi(x,0)\mathrm{d}x = 0,$$

  then in u is called a weak solution to (1). (more meaningful mathematically–motivated by multiplication by test function and integration by parts.)

It turns out the two are equivalent. (Not proven here.) Now, assume a solution that has two C^1 segments separated by a curve on which no regularity is demanded of u.

Then

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}t}\int_a^b u(x,t)\mathrm{d}x + f(u(b,t)) - f(u(a,t)) \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\left[\int_a^{x(t)} u(x,t)\mathrm{d}x + \int_{x(t)}^b u(x,t)\mathrm{d}x\right] + f(u(b,t)) - f(u(a,t)) \\
&= u(x(t^-),t)x'(t) + \int_a^{x(t)} u_t(x,t)\mathrm{d}x - u(x(t^+),t)x'(t) + \int_{x(t)}^b u_t(x,t) + f(u(b,t)) - f(u(a,t)) \\
&= u(x(t^-),t)x'(t) - \int_a^{x(t)} f(u)_x\mathrm{d}x - u(x(t^+),t)x'(t) + \int_{x(t)}^b f(u)_x\mathrm{d}x + f(u(b,t)) - f(u(a,t)) \\
&= u(x(t^-),t)x'(t) - f(u(x(t^-),t)) + f(u(a,t)) - u(x(t^+),t)x'(t) - f(u(b,t)) - f(u(x(t^+),t)) + f(u(b, t)) - f(u(a,t)) \\
&= u(x(t^-),t)x'(t) - f(u(x(t^-),t) - u(x(t^+),t)x'(t) + f(u(x(t^+),t).
\end{aligned}$$

Now use the shorthand

$$\begin{aligned} u^- &:= u(x(t^-),t) \\ u^+ &:= u(x(t^+),t) \end{aligned}$$

and write

$$0 = f(u^+) - f(u^-) - x'(t)(u^+ - u^-).$$

Now distinguish two cases:

- $u^- = u^+$: This is fine.

- $u^- \neq u^+$: We get the *Rankine-Hugoniot jump condition*:

$$x'(t) = \frac{f(u^+) - f(u^-)}{u^+ - u^-}$$

If u is piecewise C^1 and is discontinuous only along isoated curves, and if u satisfies the PDE when it is C^1, and the Rankine-Hugoniot (RH) condition along all discontinuous cruves, then u is a weak solution of (1).

Example 1. Consider the following *Riemann problem*:

$$\begin{cases} u_t + \left(\frac{u^2}{2}\right)_x = 0 \\ u(x,0) = \left\{\begin{smallmatrix} 1 & x<0, \\ -1 & x>0. \end{smallmatrix}\right. \end{cases}$$

The IC is just propagated in time to form a weak solution. (a *shock*)

Example 2. Now flip the initial conditions:

$$\begin{cases} u_t + \left(\frac{u^2}{2}\right)_x = 0 \\ u(x,0) = \left\{\begin{smallmatrix} -1 & x<0, \\ 1 & x>0. \end{smallmatrix}\right. \end{cases}$$

The propagated ICs also form a weak solution. But consider

$$u(x,t) = \begin{cases} -1 & x \leqslant -t, \\ x/t & -t < x < t, \\ 1 & x > t. \end{cases}$$

This is also a weak solution. (a *rarefaction wave*)

Oops. So, we need a third category of solutions, called *entropy solutions*, where neither uniqueness nor existence poses a big problem. Consider adding an artificial viscosity:

$$u_t^\varepsilon + f(u^\varepsilon)_x = \varepsilon u_{x,x}^\varepsilon$$

with a very small $0 < \varepsilon \ll 1$.

Then we would wish to define an entropy solution as

$$\lim_{\varepsilon \to 0} u^\varepsilon(x,t) = u(x,t)$$

in some norm. In fact, this is *the* entropy solution.

Pick a function $U(u)$ called the *entropy function* if $U''(u) \geqslant 0$, i.e. if it is convex. Then multiply the conservation law with viscosity by $U'(u^\varepsilon)$:

$$\begin{aligned} U'(u^\varepsilon)(u_t^\varepsilon + f(u^\varepsilon)_x) &= \varepsilon U'(u^\varepsilon) u_{x,x}^\varepsilon \\ U(u^\varepsilon)_t + F(u^\varepsilon)_x &= \varepsilon\big[(U'(u^\varepsilon)u_x^\varepsilon)_x - U''(u^\varepsilon)(u_x^\varepsilon)^2\big] \\ U(u^\varepsilon)_t + F(u^\varepsilon)_x &\leqslant \varepsilon(U'(u^\varepsilon)u_x^\varepsilon)_x \end{aligned}$$

where

$$F(u) = \int^u U'(v) f'(v) \mathrm{d}v \quad \Rightarrow \quad F'(u) = U'(u) f'(u).$$

To support our argument as $\varepsilon \to 0$, once again take a test function $\varphi \in C_0^2(\mathbb{R} \times \mathbb{R}^+)$, $\varphi \geqslant 0$.

$$\begin{aligned} \int_0^\infty \int_{-\infty}^\infty (U(u^\varepsilon)_t + F(u^\varepsilon)_x)\varphi \,\mathrm{d}x\,\mathrm{d}t &\leqslant \varepsilon \int_0^\infty \int_{-\infty}^\infty (U'(u^\varepsilon)u_x^\varepsilon)_x \varphi \mathrm{d}x\,\mathrm{d}t \\ \Rightarrow \int_0^\infty \int_{-\infty}^\infty U(u^\varepsilon)\varphi_t + F(u^\varepsilon)\varphi_x \,\mathrm{d}x\,\mathrm{d}t &\geqslant \varepsilon \int_0^\infty \int_{-\infty}^\infty U'(u^\varepsilon) u_x^\varepsilon \varphi_x \mathrm{d}x\,\mathrm{d}t \\ &= \varepsilon \int_0^\infty \int_{-\infty}^\infty U(u^\varepsilon)\varphi_{x,x} \mathrm{d}x\,\mathrm{d}t \end{aligned}$$

DCT allows taking the limit. We get the *entropy inequality*

$$\int_0^\infty \int_{-\infty}^\infty U(u)\varphi_t + F(u)\varphi_x \,\mathrm{d}x\,\mathrm{d}t \geqslant 0.$$

Homework #1:

- On a domain $[0, 2\pi]$, with periodic BCs, consider

  $$\begin{cases} u_t + \left(\frac{u^2}{2}\right)_x = 0 \\ u(x,0) = \frac{1}{2} + \sin x \end{cases}$$

  Find the maximum T^* such that $u(x,t) \in C^1$ for $t < T^*$.

- Write a code to solve for u when $t < T^*$. (Hint: Look for equation implicitly defining u, maybe use Newton's method). Test the code for $(0.1, 0.1)$, $(1, 0.08)$, $(\pi, 0.09)$.

Definition 3. *A conservation law is called* genuinely nonlinear *iff $f''(u) \neq 0$. If $f''(u) > 0$, it is called* convex*, if $f''(u) < 0$ it is called* concave*.*

Shocks must appear for genuinely nonlinear conservation laws under periodic or compactly supported initial conditions.

Consider a box containing the support of a test function $\varphi \in C_c^\infty(\mathbb{R} \times \mathbb{R}^+)$ and let $u(x,t)$ be piecewise C^1 with one discontinuity along $(t, x(t))$.

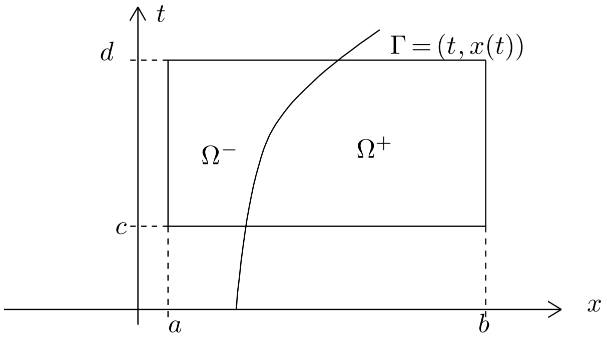


Figure 1.

Then consider

$$\begin{aligned}
0 &\leqslant -\int_c^d \int_a^b (U(u)\varphi_t + F(u)\varphi_x)\mathrm{d}x\mathrm{d}t \\
&= -\int_c^d \int_a^{x(t)} \underbrace{(U(u)\varphi_t + F(u)\varphi_x)}_{(U,F)^T \cdot \nabla\varphi}\mathrm{d}x\mathrm{d}t - \int_c^d \int_{x(t)}^b (U(u)\varphi_t + F(u)\varphi_x)\mathrm{d}x\mathrm{d}t \\
&= \int_c^d \int_a^{x(t)} \underbrace{(U(u)_t + F(u)_x)}_{=0}\varphi\mathrm{d}x\,\mathrm{d}t - \int_{\partial\Omega^-} \varphi(U(u), F(u)) \cdot \boldsymbol{n}\,\mathrm{d}s - \int_{\partial\Omega^+} \varphi(U(u), F(u)) \cdot \boldsymbol{n}\,\mathrm{d}s \\
&= \int_\Gamma \varphi \frac{x'(t)U(u^-) - F(u^-)}{\sqrt{1 + (x'(t))^2}}\,\mathrm{d}s - \int_\Gamma \varphi \frac{x'(t)U(u^+) - F(u^+)}{\sqrt{1 + x'(t)^2}}\,\mathrm{d}s \\
&= \int_\Gamma \frac{\varphi}{\sqrt{1 + x'(t)^2}}\big[x'(t)(U(u^-) - U(u^+)) - (F(u^-) - F(u^+))\big]\mathrm{d}s.
\end{aligned}$$

We obtain

$$x'(t)(U(u^-) - U(u^+)) - (F(u^-) - F(u^+)) \leqslant 0.$$

If we introduce the notation $[\![f]\!] := f(u^+) - f(u^-)$, then this condition becomes

$$x'(t)[\![U]\!] \geqslant [\![F]\!].$$

Oleinik entropy condition: For all u between u^- and u^+, we need to have

$$\frac{f(u)-f(u^-)}{u-u^-} \geqslant \underbrace{x'(t)}_{s} \geqslant \frac{f(u)-f(u^+)}{u-u^+},$$

where s is the shock speed, known from the Rankine-Hugoniot condition.

Lax's entropy condition:

$$f'(u^-) > s > f'(u^+).$$

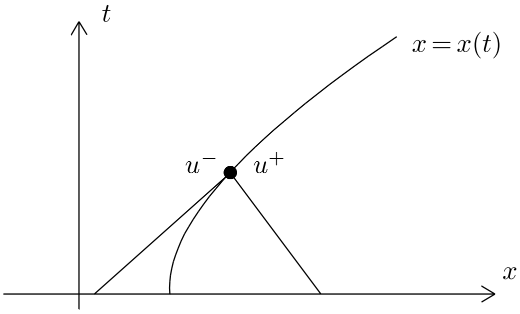


Figure 2. Illustration of Lax's entropy condition. Characteristics are going "into" shocks.

It is easy to see that the Oleinik condition implies Lax's condition. Unfortunately, the converse does not hold. Lax's entropy condition does not guarantee uniqueness–but it is a necessary condition. However, if $f''(u) \gtrless 0$ uniformly (i.e. the conservation law is genuinely nonlinear), *then* Lax's entropy condition is sufficient for u to be the entropy solution.

For $f'(u) > 0$, Lax's condition becomes even simpler. Consider

$$f'(u^-) \geqslant s = \frac{[\![f(u)]\!]}{[\![u]\!]} \geqslant f'(u^+)$$

and note that $f'(u)$ is monotonically increasing, such that the middle part is automatically satisfied. Thus, Lax's condition becomes

$$f'(u^-) \geqslant f'(u^+).$$

I.e. looking towards the right, we can only jump down.

Theorem 4. *The solutions to*

$$\begin{cases} u_t^\varepsilon + f(u^\varepsilon)_x = \varepsilon u_{x,x}^\varepsilon, \\ u^\varepsilon(x,0) = u^0(x) \end{cases}$$

are L^1-contractive. I.e. let v^ε be the solution of

$$\begin{cases} v_t^\varepsilon + f(v^\varepsilon)_x = \varepsilon v_{x,x}^\varepsilon, \\ v^\varepsilon(x,0) = v^0(x). \end{cases}$$

Then

$$\|u^\varepsilon(\,\cdot\,,t) - v^\varepsilon(\,\cdot\,,t)\|_{L^1} \leqslant \|u^0 - v^0\|_{L^1}.$$

Proof. We need to show

$$0 \geqslant \frac{\mathrm{d}}{\mathrm{d}t} \int_{-\infty}^{\infty} |u^\varepsilon(x,t) - v^\varepsilon(x,t)| \mathrm{d}x.$$

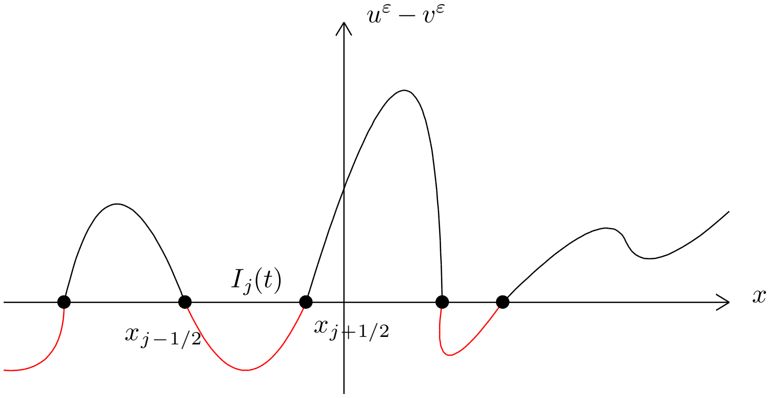


Figure 3.

Let s_j be the sign of $u^\varepsilon - v^\varepsilon$ on I_j and consider, using Leibniz's rule, the following:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{\infty}|u^\varepsilon(x,t)-v^\varepsilon(x,t)|\mathrm{d}x\\
=\ &\frac{\mathrm{d}}{\mathrm{d}t}\sum_j\int_{x_{j-1/2}}^{x_{j+1/2}}|u^\varepsilon(x,t)-v^\varepsilon(x,t)|\mathrm{d}x\\
=\ &\sum_j s_j(t)\Bigg[\underbrace{u^\varepsilon(x_{j+1/2}(t),t)-v^\varepsilon(x_{j+1/2}(t),t)}_{0}\Bigg]x'_{j+1/2}(t)\\
&-s_j(t)\Bigg[\underbrace{u^\varepsilon(x_{j+1/2}(t),t)-v^\varepsilon(x_{j+1/2}(t),t)}_{0}\Bigg]x'_{j+1/2}(t)\\
&+\int_{x_{j-1/2}}^{x_{j+1/2}}\underbrace{s_j'(t)}_{0}(u^\varepsilon(x,t)-v^\varepsilon(x,t))\mathrm{d}x\\
&+\int_{x_{j-1/2}}^{x_{j+1/2}}s_j(t)(u_t^\varepsilon(x,t)-v_t^\varepsilon(x,t))\mathrm{d}x\\
=\ &\sum_j\int_{x_{j-1/2}}^{x_{j+1/2}}s_j(t)(u_t^\varepsilon(x,t)-v_t^\varepsilon(x,t))\mathrm{d}x\\
=\ &\sum_j\int_{x_{j-1/2}}^{x_{j+1/2}}s_j(t)[-f(u^\varepsilon(x,t))_x+\varepsilon u_{x,x}^\varepsilon(x,t)+f(v^\varepsilon(x,t))_x-\varepsilon v_{x,x}^\varepsilon(x,t)]\mathrm{d}x\\
=\ &\sum_j\quad s_j(t)\Bigg\{\underbrace{-f(u^\varepsilon(x_{j+1/2}(t),t))+f(u^\varepsilon(x_{j-1/2}(t),t))+f(v^\varepsilon(x_{j+1/2}(t),t))-f(v^\varepsilon(x_{j-1/2}(t),t))}_{0}\qquad+\\
&\varepsilon\big[\textcolor{orange}{u_x^\varepsilon(x_{j+1/2}(t),t)}-\textcolor{blue}{u_x^\varepsilon(x_{j-1/2}(t),t)}-\textcolor{orange}{v_x^\varepsilon(x_{j+1/2}(t),t)}+\textcolor{blue}{v_x^\varepsilon(x_{j-1/2}(t),t)}\big]\Bigg\}\\
\leqslant\ &0.
\end{aligned}$$

To see why the orange and blue parts together each are $\geqslant 0$, just look at what's happening at the $x_{j\pm1/2}$. □

The entropy solution has a non-increasing total variation.

$$\mathrm{TV}(u):=\sup_h\int\left|\frac{u(x+h)-u(x)}{h}\right|\mathrm{d}x.$$

$$\mathrm{TV}(u(\,\cdot\,,t))\leqslant\mathrm{TV}(u^0),$$

because ...?

2 Numerics

Consider

$$\begin{cases} u_t + (\frac{u}{2})^2_x = 0 \\ u(x,0) = \left\{ \begin{smallmatrix} 1 & x<0, \\ 0 & x \geqslant 0. \end{smallmatrix} \right. \end{cases}$$

The entropy solution is

$$u(x,t) = \begin{cases} 1 & x \leqslant \frac{1}{2}t, \\ 0 & x > \frac{1}{2}t. \end{cases}$$

Note also that the analytic solution satisfies a *maximum principle*, i.e.

$$\min_x u^0(x) \leqslant u(\xi, t) \leqslant \max_x u^0(x).$$

Remember for $u_t + a\, u_x = 0$, we wrote down an *upwind scheme*:

$$u_j^{n+1} = u_j^n - a \cdot \frac{\Delta t}{\Delta x}(u_j^n - u_{j-1}^n).$$

Let's write a direct generalization, for the (equivalent...?) PDE $u_t + u\, u_x = 0$:

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} u_j^n (u_j^n - u_{j-1}^n).$$

But for $j \neq 0$, $u_j^0 - u_{j-1}^0 = 0$, and for $j = 0$, $u_j^0 = 0$. Altogether,

$$u_j^{n+1} = u_j^n.$$

Bad.

Definition 5. *A scheme to solve conservation laws is called* conservative *iff it can be written as*

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\left[\hat{f}_{j+1/2} - \hat{f}_{j-1/2}\right],$$

where $\hat{f}$ is

1. *Lipschitz continuous,*
2. *$\hat{f}(u, \cdots, u) = f(u)$* (consistency)*.*

Theorem 6. (Lax-Wendroff) *If the solution $\{u_j^n\}$ to a conservative scheme converges (as $\Delta t, \Delta x \to 0$) boundedly a.e. to a function $u(x,t)$, then u is a weak solution of the conservation law.*

Proof. Let $\varphi_j^n = \varphi(x_j, t^n)$ for $\varphi \in C_0^1$. Then

$$\begin{aligned} 0 &= \sum_n \sum_j \left(\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{\hat{f}_{j+1/2} - \hat{f}_{j-1/2}}{\Delta x} \right) \varphi_j^n \Delta x \Delta t \\ &= -\sum_n \sum_j \left(\frac{\varphi_j^n - \varphi_j^{n-1}}{\Delta t} u_j^n + \frac{\varphi_j^n - \varphi_{j-1}^n}{\Delta x} \hat{f}_{j+1/2} \right) \Delta x \Delta t \\ &\overset{\text{DCT,Conservativity}}{\longrightarrow} \int_0^\infty \int_{-\infty}^\infty (\varphi_t u + \varphi_x f(u)) \mathrm{d}x\, \mathrm{d}t = 0. \end{aligned}$$

□

Remark 7. Above, we used partial summation:

$$\sum_{j=j_1}^{j_2} a_j (b_j - b_{j-1}) = -\sum_{j=j_1}^{j_2} (a_{j+1} - a_j) b_j - a_{j_1} b_{j-1} + a_{j_2} b_{j_2}.$$

2.1 Examples of conservative schemes

2.1.1 The Godunov Scheme

The Godunov scheme for the conservation law

$$\begin{cases} u_t + f(u)_x = 0, \\ u(x,0) = u^0(x) \end{cases}$$

was derived from the fact that the Riemann problem

$$\begin{cases} u_t + f(u)_x = 0, \\ u(x,0) = \begin{cases} u_l & x < 0, \\ u_r & x \geqslant 0 \end{cases} \end{cases}$$

can be solved exactly. *Example:* (see above) For Burgers' Equation, we get

$$u(x,t) = \begin{cases} \begin{cases} u_l & x < s\,t, \\ u_r & x \geqslant s\,t, \end{cases} & u_l > u_r, \\ \begin{cases} u_l & x < u_l t, \\ x/t & u_l t \leqslant x < u_r t, \\ u_r & x \geqslant u_r t, \end{cases} & u_l < u_r, \end{cases}$$

where

$$s = \frac{f(u_r) - f(u_l)}{u_r - u_l} = \frac{\frac{1}{2}[u_r^2 - u_l^2]}{u_r - u_l} = \frac{1}{2}(u_l + u_r).$$

The same technique would work for all convex ($f''(u) > 0$) or concave conservation laws. (Also cf. book by Toro–500 pages of Riemann solutions.) Note that conservation laws have finite propagation speed. Suppose we choose a scheme where we consider the solution constant in each cell (Conceptually, imagine that this value $\bar{u}_j$ is the cell average of cell I_j–this is also how you arrive at $\bar{u}_j^0$.) If we choose Δx and Δt such that

$$\max |f'(u)| \Delta t < \Delta x,$$

then in a sequence of cells (A, B, C, D, E), then the solution in cell C in the next timestep is not influenced at all by the solution in cells A and E. Thus we only need to solve a Riemann problem at each cell interface and we're done. Then

$$\begin{aligned} \int_{t^n}^{t^{n+1}} \int_{x_{j-1/2}}^{x_{j+1/2}} (u_t + f(u)_x) \mathrm{d}x\, \mathrm{d}t &= 0 \\ \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} u^{n+1} \,\mathrm{d}x - \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} u^n \,\mathrm{d}x + \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(u_{j+1/2}) \,\mathrm{d}x - \frac{1}{\Delta x} \int_{t^n}^{t^{n+1}} f(u_{j-1/2}) \mathrm{d}x &= 0. \end{aligned}$$

Now consider that for the Riemann solution $u(x,t)$ is a function of only one variable $\xi = x/t$. In fact, the substitution

$$\begin{aligned} \bar{x} &= a\,x, \\ \bar{t} &= a\,t. \end{aligned}$$

leaves the PDE and the Riemann ICs invariant. (This is also called *self-similarity*.) Thus u is constant along $x = x_{j\pm1/2}$, making the last two integrals trivial. The Godunov scheme can then be written as

$$\bar{u}_j^1 = \bar{u}_j^0 - \frac{\Delta t}{\Delta x}\big(f(u_{j+1/2}) - f(u_{j-1/2})\big).$$

This is a conservative scheme because the flux $\hat{f}(u_j^0, u_{j+1}^0)$ depends on the right values (and Lipschitz continuity holds as well, but is a bit tricky to prove.) The numerical flux of the Godunov scheme can be written as

$$\hat{f}_{j+1/2} = \begin{cases} \min_{u_j \leqslant u \leqslant u_{j+1}} f(u) & u_j < u_{j+1}, \\ \max_{u_j \leqslant u \leqslant u_{j+1}} f(u) & u_j \geqslant u_{j+1}. \end{cases}$$

2.1.2 The Lax-Friedrichs Scheme

The numerical flux here is

$$\hat{f}_{j+1/2} = \frac{1}{2}[f(u_j) + f(u_{j+1}) - \alpha(u_{j+1} - u_j)],$$

where $\alpha = \max_u |f'(u)|$.

2.1.3 The local Lax-Friedrichs Scheme

The numerical flux here is

$$\hat{f}_{j+1/2} = \frac{1}{2}\big[f(u_j) + f(u_{j+1}) - \alpha_{j+1/2}(u_{j+1} - u_j)\big],$$

where $\alpha_{j+1/2} = \max_{(u_j, u_{j+1})} |f'(u)|$ (where we note that $(u_j,\, u_{j+1})$ is meant as a non-empty interval no matter which end of the interval is greater).

2.1.4 Roe Scheme

The numerical flux here is

$$\hat{f}_{j+1/2} = \begin{cases} f(u_j) & a_{j+1/2} \geqslant 0, \\ f(u_{j+1}) & a_{j+1/2} < 0, \end{cases}$$

where

$$a_{j+1/2} = \frac{f(u_{j+1}) - f(u_j)}{u_{j+1} - u_j}$$

is the speed of the solution as given by the RHC.

2.1.5 Engquist-Osher Scheme

The numerical flux here is

$$\hat{f}_{j+1/2} = f^+(u_j) + f^-(u_{j+1}),$$

where

$$\begin{aligned} f^+(u) &= \int_0^u \max(f'(u), 0)\mathrm{d}u + f(0), \\ f^-(u) &= \int_0^u \min(f'(u), 0)\mathrm{d}u. \end{aligned}$$

2.1.6 Lax-Wendroff Scheme

Consider

$$\begin{aligned} u_t &= -f(u)_x \\ u_{t,t} &= -f(u)_{x,t} = -(f(u)_t)_x = -(f'(u)u_t)_x = (f'(u)f(u)_x)_x. \end{aligned}$$

The general idea is:

- Repeatedly replace time by space derivatives by using the PDE,
- Discretize space derivatives by (2nd order central) FD formulae.

Derivation:

$$\begin{aligned} u^{n+1} &= u^n + \Delta t\, u_t^n + \frac{\Delta t^2}{2} u_{t,t}^n \\ &= u^n - \Delta t\, f(u^n)_x + \frac{\Delta t^2}{2}(f'(u)f(u)_x)_x \\ u_j^{n+1} &= u_j^n - \Delta t \frac{f(u_{j+1}^n) - f(u_{j-1}^n)}{2\Delta x} + \frac{\Delta t^2}{2}\left[f'(u_{j+1/2}^n)\frac{f(u_{j+1}^n) - f(u_j^n)}{\Delta x} - \right. \\ &\quad \left. f'(u_{j-1/2}^n)\frac{f(u_j^n - f(u_{j-1}^n)}{\Delta x}\right] \Big/ \Delta x, \end{aligned}$$

where

$$u_{j+1/2}^n = \frac{u_j^n + u_{j+1}^n}{2}.$$

The numerical flux becomes

$$\hat{f}_{j+1/2} = \frac{1}{2}\big[f(u_j) + f(u_{j+1}) - \lambda f'(u_{j+1/2})(f(u_{j+1}) - f(u_j))\big],$$

where

$$\lambda = \frac{\Delta t}{\Delta x}.$$

2.1.7 MacCormack Scheme

The idea behind MacCormack is of the "predictor-corrector" sort.

$$\begin{aligned} u_j^{n+1/2} &= u_j^n - \lambda(f(u_j^n) - f(u_{j-1}^n)), \\ u_j^{n+1} &= \frac{1}{2}\Big[u_j^n + u_j^{n+1/2} + \lambda\Big[f(u_{j+1}^{n+1/2}) - f(u_j^{n+1/2})\Big]. \end{aligned}$$

The numerical flux is a bit ugly:

$$\hat{f}_{j+1/2} = \frac{1}{2}[f(u_j) + f(u_j - \lambda(f(u_j) - f(u_{j-1})))].$$

Homework #2:

1. Code the Godunov and Lax-Friedrichs scheme for solving a Riemann problem of Burgers' Equation. Test the code with

   a) $u_l = 1$, $u_r = -0.5$.

   b) $u_l = -0.5$, $u_r = 1$

   using $N = 160$ points equally spaced. Show the solution graphically along with the exact solution.

2. Find the formula for the entropy solution of

   $$\begin{cases} u_t + f(u)_x = 0, \\ u(x,0) = \left\{\begin{smallmatrix} u_l & x<0, \\ u_r & x>0 \end{smallmatrix}\right. \end{cases}$$

   where $f''(u) > 0$.

3. Show that the Godunov flux and the Roe flux are both Lipschitz-continuous.

Definition 8. *A scheme*

$$\begin{aligned} u_j^{n+1} &= u_j^n - \lambda(\hat{f}(u_{j-p}, \ldots, u_{j+q}) - \hat{f}(u_{j-p-1}, \ldots, u_{j+q-1})) \\ &\equiv G(u_{j-p-1}, \ldots, u_{j+q}) \end{aligned}$$

is called a montone scheme *if G is a monotonically nondecreasing function $G(\uparrow, \uparrow, \ldots, \uparrow)$ of each argument.*

In the special case of 3-point schemes

$$\hat{f}(u_j, u_{j+1})$$

the scheme is a monotone if $f(\uparrow, \downarrow)$ plus a restriction on λ:

$$G(u_{j-1}, u_j, u_{j+1}) = u_j - \lambda[\hat{f}(u_j, u_{j+1}) - \hat{f}(u_{j-1}, u_j)].$$

Clearly, if $\hat{f}(\uparrow, \downarrow)$, then $G(\uparrow, ?, \uparrow)$. To clean up the second argument, consider

$$\frac{\partial G}{\partial u_j} = 1 - \lambda[\underbrace{\hat{f}_1 - \hat{f}_2}_{\geqslant 0}] \geqslant 0.$$

If $\lambda(\hat{f}_1 - \hat{f}_2) \leqslant 1$, then $G(\uparrow, \uparrow, \uparrow)$.

Examples: The Lax-Friedrichs flux is monotone:

$$\begin{aligned} \hat{f}^{\mathrm{LF}}(u_j, u_{j+1}) &= \frac{1}{2}[f(u_j) + f(u_{j-1}) - \alpha(u_{j+1} - u_j)] \quad \text{for} \quad \alpha = \max_u |f'(u)|, \\ \hat{f}_1^{\mathrm{LF}} &= \frac{1}{2}[f'(u_j) + \alpha] \geqslant 0, \\ \hat{f}_2^{\mathrm{LF}} &= \frac{1}{2}[f'(u_{j+1}) + \alpha] \leqslant 0. \end{aligned}$$

Theorem 9. *Good properties of monotone schemes:*

1. *$u_j \leqslant v_j$ for all j ("$u \leqslant v$") implies $G(u)_j \leqslant G(v)_j$ for all j.*
2. Local maximum principle:
$$\min_{i \in \text{stencil around } j} u_i \leqslant G(u)_j \leqslant \max_{i \in \text{stencil around } j} u_i.$$
3. *L^1-contraction: (this was already obtained for the PDE)*
$$\|G(u) - G(v)\|_{L^1} \leqslant \|u - v\|.$$
4. *This immediately implies the* T*otal* V*ariation* D*iminishing (TVD) property:*
$$\|G(u)\|_{\mathrm{BV}} \leqslant \|u\|_{\mathrm{BV}}.$$

Proof. 1 is just the definition.

2. Fix j. Take
$$v_i = \begin{cases} \max_{k \in \text{stencil arond } i} u_k & \text{if } i \in \text{stencil around } j, \\ u_i & \text{otherwise.} \end{cases}$$

Then clearly $u_i \leqslant v_i$ for all i, so that
$$G(u)_j \leqslant G(v)_j = v_j = \max_{i \in \text{stencil around } j} u_i.$$

Other way around runs in an analogous fashion.

3. Define
$$a \vee b = \max(a, b), \quad a \wedge b = \min(a, b), \quad a^+ = a \wedge 0, \quad a^- = a \vee 0.$$

Then let
$$w_j := u_j \vee v_j = v_j + (u_j - v_j)^+. \quad (*)$$

We have
$$G(u)_j \leqslant G(w)_j \geqslant G(v)_j \quad \forall j$$

by property 1. Then
$$G(w)_j - G(v)_j \geqslant \begin{cases} 0 & \forall j, \\ G(u_j) - G(v_j) & \forall j. \end{cases}$$

Thus
$$G(w)_j - G(v)_j \geqslant (G(u)_j - G(v)_j)^+.$$

Therefore
$$\sum_j (G(u)_j - G(v_j))^+ \leqslant \sum_j (G(w)_j - G(v))_j \overset{(**)}{=} \sum_j w_j - v_j \overset{(*)}{=} \sum_j (u_j - v_j)^+.$$

because we are treating a *conservation* law, meaning
$$\sum_j u_j^{n+1} = \sum_j u_j^n, \quad (**)$$

which holds for *conservative schemes*. (Why?) Also consider
$$\begin{aligned} \sum_j |G(u)_j - G(v)_j| &= \sum_j (G(u)_j - G(v)_j)^+ + \sum_j (G(u)_j - G(v)_j)^- \\ &\leqslant \sum_j (u_j - v_j)^+ + \sum_j (v_j - u_j)^+ \\ &= \sum_j |u_j - v_j|. \end{aligned}$$

(This is also called the *Crandall-Tartar lemma.*)

4: Take $v_j = u_{j+1}$ in 3. □

Theorem 10. *Solutions to monotone schemes satisfy* all *entropy conditions.*

Proof. We'll prove a particular case, namely

$$U(u) = |u - c|$$

for any $c \in \mathbb{R}$. Then

$$U'(u) = \begin{cases} -1 & u < c, \\ 1 & u > c \end{cases}$$

and $U''(u) = 2\delta(x - c) \geqslant 0$.

> (Recall that entropy conditions were of the form, "pick an entropy function $U''(u) \geqslant 0$, then $U(u)_t + F(u)_x = 0$", where F is the entropy flux
>
> $$F(u) = \int_c^u U'(u) f'(u) \mathrm{d}u$$
>
> satisfying $F'(u) = U'(u) f'(u)$.)

Here we let

$$F(u) = \operatorname{sign}(u - c)(f(u) - f(c)).$$

We claim that the *cell entropy inequality* is true, i.e.

$$\frac{U(u_j^{n+1}) - U(u_j^n)}{\Delta t} + \frac{\hat{F}_{j+1/2} - \hat{F}_{j-1/2}}{\Delta x} \leqslant 0,$$

where

$$\hat{F} = \hat{f}(c \vee u) - \hat{f}(c \wedge u).$$

Observe that we've abused notation a bit, i.e.

$$\hat{f}(\alpha) := \hat{f}(\alpha, \alpha, \ldots, \alpha).$$

First step: Try to show

$$|u_j^n - c| - \lambda(\hat{F}_{j+1/2} - \hat{F}_{j-1/2}) \;=\; G(c \vee u)_j - G(c \wedge u)_j.$$

Now consider:

$$\begin{aligned} \text{I:} \quad G(c \vee u)_j &= (c \vee u_j) - \lambda(\hat{f}(c \vee u)_{j+1/2}) - \hat{f}(c \vee u_{j-1/2})) \\ \text{II:} \quad G(c \wedge u)_j &= (c \wedge u_j) - \lambda(\hat{f}(c \wedge u)_{j+1/2}) - \hat{f}(c \wedge u_{j-1/2})) \\ \text{I} - \text{II:} \quad 0 \leqslant G(c \vee u)_j - G(c \wedge u)_j &= |u_j - c| - \lambda(\hat{F}_{j+1/2} - \hat{F}_j). \end{aligned}$$

Next, note that

$$\begin{aligned} c &\overset{*}{=} G(c, \ldots, c) \leqslant G(c \vee u)_j, \\ u_j^{n+1} &= G(u^n)_j \leqslant G(c \vee u)_j, \\ \Rightarrow c \vee u_j^{n+1} &\leqslant G(c \vee u^n)_j, \end{aligned}$$

where the step " $*$ " is true because if the arguments of G are constant, then only the u_j^n term comes into play, just yielding back the argument.

Also

$$-c \vee u_j^{n+1} \leqslant -G(c \wedge u^n)_j.$$

Then

$$\begin{aligned} U(u_j^{n+1}) &= |u_j^{n+1} - c| \leqslant G(c \vee u^n)_j - G(c \wedge u^n)_j \\ &= \underbrace{|u_j^n - c|}_{U(u_j^n)} - \lambda(\hat{F}_{j+1/2} - \hat{F}_{j-1/2}). \end{aligned}$$

□

Theorem 11. (Godunov) *Monotone schemes are at most first-order accurate.*

After this depressing result, we will have to look for different classes of schemes. For example, in order of decreasing strength:

- Monotone: see above.
- TVD: A scheme is TVD if
$$\mathrm{TV}(u^{n+1}) \leqslant \mathrm{TV}(u^n).$$
- Monotonicity-preserving: A scheme is monotonicity-perserving if
$$\{u_{j+1}^n \geqslant u_j^n \forall j\} \Rightarrow \{u_{j+1}^{n+1} \geqslant u_j^{n+1} \forall j\}.$$

Let's prove that the above is actually in order of decreasing strength, i.e.

Theorem 12. *A TVD scheme is monotonicity-preserving.*

Proof. Assume $u_{j+1}^n \geqslant u_j^n$ for all j. If there exists a j_0 such that $u_{j_0+1}^{n+1} < u_{j_0}^{n+1}$. Modify u to be constant outside the stencil used to compute $u_{j_0}^{n+1}$ and $u_{j_0+1}^{n+1}$. But the reversal of the order of these two values means that the TVD property is violated. □

Later in this class, a theorem by Godunov will show that all the above properties are actually the same, and thus first-order, and thus useless. :-/

Definition 13. *A scheme is called a "*linear scheme*" if it is linear when applied to a linear PDE:*
$$u_t + a\, u_x = 0,$$
where a is a constant.

A linear scheme for
$$u_t + u_x = 0 \tag{5}$$
can be written as
$$u_j^{n+1} = \sum_{l=-k}^{k} c_l(\lambda) u_{j-l}^n,$$
where $c_l(\lambda)$ are constants which may depend on $\lambda = \Delta t / \Delta x$. A linear scheme for (5) is monotone iff
$$c_l(\lambda) \geqslant 0 \quad \forall l.$$
This is why they are also called "*positive schemes*".

Theorem 14. *For linear schemes, monotonicity-preserving ⇒ monotone.*

Corollary 15. *For linear schemes, monotonicity-preserving and TVD schemes are at most first order accurate.*

Proof. (of Theorem 14) If the above linear scheme is monotonicity-perserving, then consider
$$u_i = \begin{cases} 0 & i \leqslant -\alpha, \\ 1 & i > -\alpha. \end{cases}$$
This is a monotone function. Then
$$\begin{aligned} \text{(I)} \quad u_{j+1}^{n+1} &= \sum_{l=-k}^{k} c_l(\lambda) u_{j+1}^n \\ \text{(II)} \quad u_j^{n+1} &= \sum_{l=-k}^{k} c_l(\lambda) u_{j-l}^n \\ \text{(I)} - \text{(II)}: \quad \Delta u_j^{n+1} &= \sum_{l=-k}^{k} c_l(\lambda) \Delta u_{j-l}^n \end{aligned}$$

where we note that $\Delta u_\alpha^n = 1$ if $m = -\alpha$, and zero otherwise.

$$\Delta u_0^{n+1} = \sum_{l=-k}^{k} c_l(\lambda)\Delta u_{-l}^n = c_\alpha(\lambda) \geqslant 0,$$

due to the requirement of monotonicty-preserving-ness, meaning all $c_\alpha(\lambda) \geqslant 0$, such that the scheme is monotone. □

So, we have

$$\text{monotonicity-preserving (MP)} \overset{*}{\Rightarrow} \text{monotone} \Rightarrow \text{TVD} \Rightarrow \text{MP}$$

where the implication "$*$" only holds for linear schemes.

For a scheme to be consistent, $\tau_j^n = 0$ if u is a constant solution (where τ_j^n is the local truncation error). For a scheme to be at least first order accurate, $\tau_j^n = 0$ if u is a linear solution of the PDE.

Consider a linear scheme

$$u_j^{n+1} = \sum_l c_l u_{j-l}^n.$$

Plug a constant in there, and we obtain

$$1 = \sum_l c_l.$$

Plug a linear term in there, and obtain

$$\begin{aligned} j\Delta x - (n+1)\Delta t &= \sum_l c_l((j-l)\Delta x - n\Delta t) \\ -\Delta t &= \Delta x \sum_l (-l)c_l \\ \sum_l l\, c_l &= \lambda \end{aligned}$$

For a quadratic term, we would get

$$\sum_l l^2 c_l = \lambda^2.$$

So, now try to derive a contradiction between any two of the above to refute second-order. To that end, define

$$\boldsymbol{a} = (l\sqrt{c_l})_{l=-k}^k, \quad \boldsymbol{b} = (\sqrt{c_l})_{l=-k}^k$$

and now use Cauchy-Schwarz:

$$\lambda^2 = |\boldsymbol{a}\cdot\boldsymbol{b}|^2 \overset{*}{\leqslant} \left(\sum_l l^2 c_l\right)\left(\sum_l c_l\right) = \lambda^2,$$

where equality in "$*$" holds only if $\boldsymbol{a}$ and $\boldsymbol{b}$ are linearly dependent, i.e.

$$l\sqrt{c_l} = \alpha\sqrt{c_l},$$

where α is just some constant independent of l.

Theorem 16. (Godunov) *A linear monotone (TVD) scheme is at most first-order accurate.*

2.2 Higher-order TVD Schemes

Consider

$$u_t + f(u)_x = 0,$$

where we will worry about the computation of the spatial derivative now and about the time derivative later. Then we can use backward differences

$$\frac{f(u_j) - f(u_{j-1})}{\Delta x}$$

for first-order accuracy or

$$\frac{f(u_{j+1}) - f(u_{j-1})}{2\Delta x}$$

for second-order accuracy or

$$\frac{\frac{3}{2}f(u_j) - 2f(u_{j-1}) + \frac{1}{2}f(u_{j-2})}{\Delta x}$$

for third-order.

2.2.1 General Framework of a Conservative Finite-Volume Scheme

Consider our conventional notation of $I_j = [x_{j-1/2}, x_{j+1/2}]$, where $\Delta x_j = x_{j+1/2} - x_{j-1/2}$. Now integrate the PDE:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{x_{j-1/2}}^{x_{j+1/2}} u\,\mathrm{d}x + f(u(x_{j+1/2})) - f(u(x_{j-1/2})) = 0$$

Denote

$$\bar{u}_j = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} u\,\mathrm{d}x.$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{u}_j + \frac{1}{\Delta x_j}[f(u(x_{j+1/2}, t)) - f(u(x_{j-1/2}, t)).$$

A finite volume scheme is of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{u}_j + \frac{1}{\Delta x_j}\Big[\hat{f}_{j+1/2} - \hat{f}_{j-1/2}\Big],$$

where $\hat{f}_{j+1/2}$ is the numerical flux. We want

$$\hat{f}_{j+1/2} \approx f(u(x_{j+1/2}, t)).$$

For the time being, let's assume $f'(u) \geqslant 0$ and $\hat{f}_{j+1/2} = f(\bar{u}_j)$, which is the numerical flux for Godunov, Roe, Engquist-Osher. See below for the case of unknown sign.

$$\hat{f}_{j+1/2} = \hat{f}(\bar{u}_j, \bar{u}_{j+1}),$$

where $\hat{f}(\uparrow, \downarrow)$. So, we can try to compute $u_{j+1/2}$ using the information $\{\bar{u}_j, \bar{u}_{j+1}\}$ as

$$\begin{aligned} u_{j+1/2}^{(1)} &= \frac{1}{2}(\bar{u}_j + \bar{u}_{j+1}), \\ u_{j+1/2}^{(2)} &= \frac{3}{2}\bar{u}_j - \frac{1}{2}\bar{u}_{j-1}, \end{aligned}$$

so that

$$\begin{aligned} \hat{f}_{j+1/2}^{(1)} &= f(u_{j+1/2}^{(1)}) = f\left(\frac{1}{2}(\bar{u}_j + \bar{u}_{j+1})\right), \\ \hat{f}_{j+1/2}^{(2)} &= f(u_{j+1/2}^{(2)}) = f\left(\frac{1}{2}(3\bar{u}_j - \bar{u}_{j-1})\right). \end{aligned}$$

The above fluxes are 2nd order accurate, and are called the 2nd order central and upwind flux, respectively. ($u^{(1)}$ is gained from the line connecting the cell centers at the cell averages of I_j and I_{j+1}. $u^{(2)}$ is the same for I_j and I_{j-1}.)

The step from $\{\bar{u}_j\} \rightarrow \{u_{j+1/2}\}$ is called *reconstruction*.

$$\begin{aligned} \hat{f}_{j+1/2}^{(1)} &= f\left(\bar{u}_j + \underbrace{\frac{1}{2}(\bar{u}_{j+1} - \bar{u}_j)}_{\tilde{u}_j^{(1)}}\right), \\ \hat{f}_{j-1/2}^{(2)} &= f\left(\bar{u}_j + \underbrace{\frac{1}{2}(\bar{u}_j - \bar{u}_{j-1})}_{\tilde{u}_j^{(2)}}\right). \end{aligned}$$

$\tilde{u}_j^{(i)}$ measures the distance from the cell average $\bar{u}_j$ to $u_{j+1/2}^{(1)}$. Now define

$$\operatorname{minmod}(a,b) := \begin{cases} a & |a| < |b|, a\,b > 0, \\ b & |b| < |a|, a\,b > 0, \\ 0 & a\,b \leqslant 0 \end{cases}$$

and set

$$\tilde{u}_j := \operatorname{minmod}\left(\tilde{u}_j^{(1)}, \tilde{u}_j^{(2)}\right).$$

Then consider

$$\hat{f}_{j+1/2}^{(3)} = f(\bar{u}_j + \tilde{u}_j).$$

Lemma 17. (Harten) *If a scheme can be written as*

$$\bar{u}_{j+1} = \bar{u}_j + \lambda\left(C_{j+1/2}\Delta_+\bar{u}_j - D_{j-1/2}\Delta_-\bar{u}_j\right)$$

with $C_{j+1/2} \geqslant 0$, $D_{j+1/2} \geqslant 0$, $1 - \lambda(C_{j+1/2} + D_{j+1/2}) \geqslant 0$ and $\lambda = \Delta t/\Delta x$, then it is TVD. As a matter of notation, we have

$$\begin{aligned} \Delta_+ u_j &= u_{j+1} - u_j, \\ \Delta_- u_j &= u_j - u_{j-1}. \end{aligned}$$

Proof. Write

$$\begin{aligned} \Delta_+\bar{u}_j^{n+1} &= \Delta_+\bar{u}_j^n + \lambda(C_{j+3/2}\Delta_+\bar{u}_{j+1}^n - D_{j+1/2}\Delta_+\bar{u}_j^n - C_{j+1/2}\Delta_+\bar{u}_j^n + D_{j-1/2}\Delta_-\bar{u}_j^n) \\ &= [1 - \lambda(C_{j+1/2} + D_{j+1/2})]\Delta_+\bar{u}_j^n + \lambda C_{j+3/2}\Delta_+\bar{u}_{j+1}^n + \lambda D_{j-1/2}\Delta_-\bar{u}_j^n. \end{aligned}$$

Thus

$$\begin{aligned} |\Delta_+\bar{u}_j^{n-1}| &\leqslant \left[1 - \lambda(C_{j+1/2} + D_{j+1/2})\right]|\Delta_+\bar{u}_j^n| + \lambda\underbrace{C_{j+3/2}|\Delta + \bar{u}_{j+1}^n|}_{C_{j'+1/2}|\Delta_+\bar{u}_{j'}^n|} + \lambda\underbrace{D_{j-1/2}|\Delta_-\bar{u}_j^n|}_{D_{j''+1/2}|\Delta_+\bar{u}_{j''}^n|}. \\ \sum_j |\Delta_+\bar{u}_j^{n-1}| &\leqslant \sum_j \left[1 - \lambda(C_{j+1/2} + D_{j+1/2}) + \lambda C_{j+1/2} + \lambda D_{j+1/2}\right]|\Delta_+\bar{u}_j^n| \\ \mathrm{TV}(\bar{u}_j^{n+1}) &\leqslant \mathrm{TV}(u_j^n), \end{aligned}$$

which proves the claim. □

Next, prove that the scheme we designed above is TVD using Harten's Lemma. Rewrite

$$\bar{u}_j^{n+1} = \bar{u}_j - \lambda[f(\bar{u}_j + \tilde{u}_j) - f(\bar{u}_{j-1} + \tilde{u}_{j-1})] = \bar{u}_j - \lambda\left[-D_{j-1/2}\Delta_-\bar{u}_j\right],$$

with

$$\begin{aligned} D_{j-1/2} &= \frac{f(\bar{u}_j + \tilde{u}_j) - f(\bar{u}_{j-1} + \tilde{u}_{j-1})}{\bar{u}_j - \bar{u}_{j-1}} = f'(\xi)\frac{\bar{u}_j - \bar{u}_{j-1} + \tilde{u}_j - \tilde{u}_{j-1}}{\bar{u}_j - \bar{u}_{j-1}} \\ &= f'(\xi)\left[1 + \underbrace{\frac{\tilde{u}_j}{\bar{u}_j - \bar{u}_{j-1}}}_{0 \leqslant \cdot \leqslant \frac{1}{2}} - \underbrace{\frac{\tilde{u}_{j-1}}{\bar{u}_j - \bar{u}_{j-1}}}_{0 \leqslant \cdot \leqslant \frac{1}{2}}\right] \geqslant 0 \end{aligned}$$

Thus our scheme is TVD. □

We also get a condition for the CFL number.

$$D_{j-1/2} \leqslant 3/2 f'(\xi) \leqslant \frac{3}{2}\max|f'(\xi)|,$$

which comes from

$$1 - \lambda D_{j-1/2} \geqslant 1 - \frac{3}{2}\lambda\max|f'(\xi)| \geqslant 0 \Leftarrow \boxed{\lambda\max|f'(\xi)| \leqslant \frac{2}{3}.}$$

If we use a 2nd order Runge-Kutta method like

$$\begin{aligned} \bar{u}^{(1)} &= L(\bar{u}^n), \\ \bar{u}^{n+1} &= \frac{1}{2}\left(\bar{u}^n + L(\bar{u}^{(1)})\right), \end{aligned}$$

then

$$\begin{aligned}\mathrm{TV}(\bar{u}^{(1)}) &\leqslant \mathrm{TV}(\bar{u}^{n})\\ \mathrm{TV}(\bar{u}^{n+1}) &\leqslant \frac{1}{2}\mathrm{TV}(\bar{u}^{n})+\frac{1}{2}\mathrm{TV}(L(\bar{u}^{(1)}))\\ &\leqslant \frac{1}{2}\mathrm{TV}(\bar{u}^{n})+\frac{1}{2}\mathrm{TV}(\bar{u}^{(1)})\\ &\leqslant \frac{1}{2}\mathrm{TV}(\bar{u}^{n})+\frac{1}{2}\mathrm{TV}(\bar{u}^{n})\\ &= \mathrm{TV}(\bar{u}^{n}).\end{aligned}$$

The scheme treated here is called *MUSCL* (“monotone upstream scheme for conservation laws”).

Homework #3:

1. Prove: Conservative montone schemes are at most first order accurate.
2. Prove: For every convex entropy

   $$U''(u)\geqslant 0$$

   and a conservative monotone scheme, there exists a consistent ($\hat{F}(u, \ldots, u) = \hat{F}(u)$) entropy flux $\hat{F}_{j+1/2}$ such that the following cell entropy inequality holds

   $$\frac{U(u_j^{n+1})-U(u_j^n)}{\Delta t}+\frac{\hat{F}_{j+1/2}-\hat{F}_{j-1/2}}{\Delta x}\leqslant 0,$$

   where

   $$u_j^{n+1}=u_j^n-\lambda(\hat{f}_{j+1/2}-\hat{f}_{j-1/2})=H(\underset{\uparrow}{u_{j-p}^n},\ldots,\underset{\uparrow}{u_{j+q}^n}).$$

   (We proved this for $U(u)=|u-c|$.)
3. Code:

   $$\begin{cases}u_t+\left(\frac{u^2}{2}\right)_x=0\\ u(x,0)=1+\frac{1}{2}\sin(x)\end{cases}$$

   on $0\leqslant x\leqslant 2\pi$ to (i) $t=1.0$ and (ii) $t=3.0$. Use a uniform grid with $N=20,40,80,160,320$. Use

   i. First order Godunov (upwinding)
   ii. 2nd order central ($\tilde{u}^{(1)}$)
   iii. 2nd order upwind ($\tilde{u}^{(2)}$)
   iv. MUSCL (minmod)

   For (i): tables of L^1 errors and orders. For (ii): Figures for $N=40$.

2.2.2 Generalized MUSCL Scheme

We are still considering

$$u_t+f(u)_x=0,$$

with a scheme of the form

$$\bar{u}_j^{n+1}=\bar{u}_j^n-\lambda[\hat{f}(u_{j+1/2}^-,u_{j+1/2}^+)-\hat{f}(u_{j-1/2}^-,u_{j-1/2}^+)],$$

where $\hat{f}(\uparrow,\downarrow)$ is a monotone flux. Before we can seriously start considering the above scheme, we need to specify the reconstruction step, which achieves the mapping

$$\{\bar{u}_j\}\mapsto\{\bar{u}_{j+1/2}^{\pm}\}.$$

Procedure:

From $\{\bar{u}_j\}$, we obtain the reconstructed functions $P_j(x)$ defined on $I_j = (x_{j-1/2}, x_{j+1/2})$ and then take $u_{j+1/2}^{-} = P_j(x_{j+1/2})$, $u_{j+1/2}^{+} = P_{j+1}(x_{j+1/2})$. Conditions on P_j:

- $\frac{1}{\Delta x}\int_{I_j} P_j(x)\mathrm{d}x = \bar{u}_j$,
- $\frac{1}{\Delta x}\int_{I_{j+l}} P_j(x)\mathrm{d}x = \bar{u}_{j+l}$ for some set of $l \neq 0$. (accuracy)

3rd order reconstruction formulas:

$$\begin{aligned} u_{j+1/2}^{(1)} &= \frac{1}{3}\bar{u}_{j-2} - \frac{7}{6}\bar{u}_{j-1} + \frac{11}{6}\bar{u}_j, \\ u_{j+1/2}^{(2)} &= -\frac{1}{6}\bar{u}_{j-1} + \frac{5}{6}\bar{u}_j - \frac{1}{3}\bar{u}_{j+1}. \\ u_{j+1/2}^{(3)} &= \frac{1}{3}\bar{u}_j + \frac{5}{6}\bar{u}_{j+1} - \frac{1}{6}\bar{u}_{j+2}. \end{aligned}$$

We could then choose

$$u_{j+1/2}^{-} = u_{j+1/2}^{(2)}, \quad u_{j+1/2}^{+} = u_{j+1/2}^{(3)}$$

and once more obtain a linear scheme, which is third order accurate and, by Godunov's theorem, should be oscillatory. Now define

$$\begin{aligned} \tilde{u}_j &= u_{j+1/2}^{-} - \bar{u}_j, \\ \tilde{\tilde{u}}_{j+1} &= u_{j+1/2}^{+} + \bar{u}_{j+1}, \end{aligned}$$

or equivalently

$$\begin{aligned} u_{j+1/2}^{-} &= \bar{u}_j + \tilde{u}_j, \\ u_{j+1/2}^{+} &= \bar{u}_{j+1} - \tilde{\tilde{u}}_{j+1}. \end{aligned}$$

Then, remember our previous modification of the reconstruction and do something analogous:

$$\begin{aligned} \tilde{u}_j^{\text{mod}} &= \operatorname{minmod}(\tilde{u}_j, \bar{u}_{j+1} - \bar{u}_j, \bar{u}_j - \bar{u}_{j-1}), \\ \tilde{\tilde{u}}_j^{\text{mod}} &= \operatorname{minmod}(\tilde{\tilde{u}}_j, \bar{u}_{j+1} - \bar{u}_j, \bar{u}_j - \bar{u}_{j-1}) \end{aligned}$$

and with that

$$\begin{aligned} u_{j+1/2}^{-,\text{mod}} &= \bar{u}_j + \tilde{u}_j^{\text{mod}} \\ u_{j+1/2}^{+,\text{mod}} &= \bar{u}_j - \tilde{\tilde{u}}_{j+1}^{\text{mod}}. \end{aligned}$$

To show that this modification does not destroy much accuracy and is in fact TVD, consider

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \lambda[\underbrace{\hat{f}(u_{j+1/2}^{-,\text{mod}}, u_{j+1/2}^{+,\text{mod}}) - \hat{f}(u_{j+1/2}^{-,\text{mod}}, u_{j-1/2}^{+,\text{mod}})}_{(2)} + \underbrace{\hat{f}(u_{j+1/2}^{-,\text{mod}}, u_{j-1/2}^{+,\text{mod}}) - \hat{f}(u_{j-1/2}^{-,\text{mod}}, u_{j-1/2}^{+,\text{mod}})}_{(1)}],$$

where these terms correspond to the marked terms in the assumption of Harten's lemma:

$$\bar{u}_{j+1} = \bar{u}_j + \lambda\left(\underbrace{C_{j+1/2}\Delta_+\bar{u}_j}_{(2)} - \underbrace{D_{j-1/2}\Delta_-\bar{u}_j}_{(1)}\right).$$

Now consider

$$\begin{aligned} D_{j-1/2} &= \frac{\hat{f}(u_{j+1/2}^{-,\text{mod}}, u_{j-1/2}^{+,\text{mod}}) - \hat{f}(u_{j-1/2}^{-,\text{mod}}, u_{j-1/2}^{+,\text{mod}})}{\bar{u}_j - \bar{u}_{j-1}} \\ &= \hat{f}_1(\xi, u_{j-1/2}^{+,\text{mod}})\,\frac{\bar{u}_j + \tilde{u}_j^{\text{mod}} - \bar{u}_{j-1} - \tilde{u}_{j-1}^{\text{mod}}}{\bar{u}_j - \bar{u}_{j-1}} \\ &= \underbrace{\hat{f}_1(\xi, u_{j-1/2}^{+,\text{mod}})}_{\geqslant 0\,(\text{monotonicity})} \cdot \underbrace{\left[1 + \underbrace{\frac{\tilde{u}_j^{\text{mod}}}{\bar{u}_j - \bar{u}_{j-1}}}_{0\leqslant\cdot\leqslant 1} - \underbrace{\frac{\tilde{u}_{j-1}^{\text{mod}}}{\bar{u}_j - \bar{u}_{j-1}}}_{0\leqslant\cdot\leqslant 1}\right]}_{0\leqslant\cdot\leqslant 2}. \\ &\geqslant 0. \end{aligned}$$

Claim:

> In smooth and monotone regions the scheme maintains its original high order accuracy.

Consider the following Taylor expansions:

$$\begin{aligned}
u_{j+1/2}^{-} &= u(x_{j+1/2}) + O(\Delta x^r), \quad r \geqslant 2\\
&= u(x_j) + u_x(x_j)\frac{\Delta x}{2} + O(\Delta x^2).\\
\bar{u}_j &= \frac{1}{\Delta x}\int_{I_j} u(x)\mathrm{d}x\\
&= \frac{1}{\Delta x}\int_{I_j}\left[u(x_j) + u_x(x-x_j) + u_{x,x}\frac{(x-x_j)^2}{2} + O(\Delta x^3)\right]\mathrm{d}x\\
&= u(x_j) + O(\Delta x^2).\\
\tilde{u}_j &= u_{j+1/2}^{-} - \bar{u}_j\\
&= u_x(x_j)\frac{\Delta x}{2} + O(\Delta x^2).\\
\bar{u}_{j+1} - \bar{u}_j &= u(x_{j+1}) - u(x_j) + O(\Delta x^2)\\
&= u_x(x_j)\Delta x + O(\Delta x^2)\\
\bar{u}_j - \bar{u}_{j-1} &= u_x(x_j)\Delta x + O(\Delta x^2).
\end{aligned}$$

Observe that the second and third arguments of the minmod function–it is about half as big as the first one. The monotonicity assumption above has the consequence that we may neglect the second-order terms in favor of the first-order one.

Theorem 18. (Osher) *TVD schemes are at most first-order accurate near smooth extrema.*

A simple argument by Harten shows something similar. Why are we restricted near smooth extrema? Suppose we are considering $u_t + u_x = 0$.

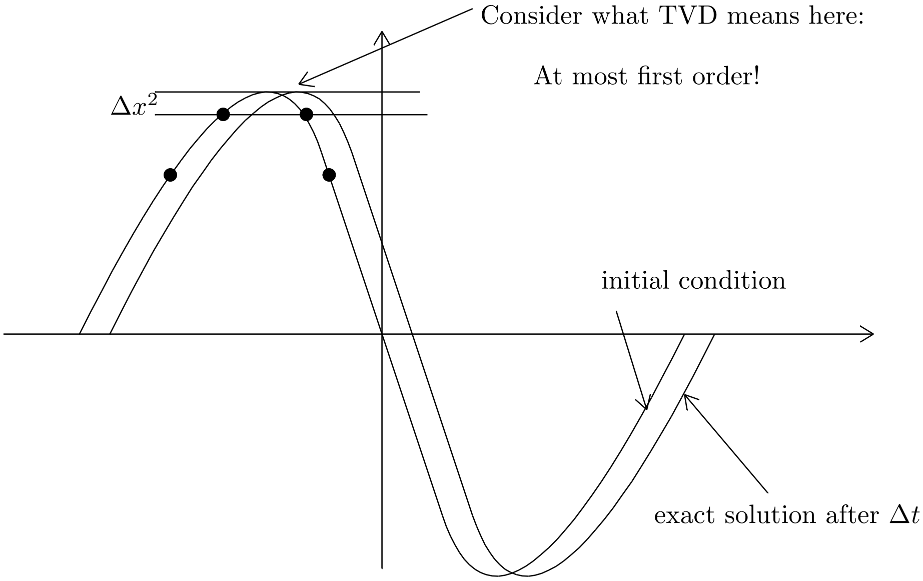

Consider what TVD means here:
At most first order!
Δx^2
initial condition
exact solution after Δt

Figure 4. Why TVD schemes don't do so well near smooth extrema.

What routes can we take out of this dilemma? Relax TVD: Only demand TVB.

$$\mathrm{TV}(\bar{u}^{n+1}) \leqslant (1 + C\Delta t)\mathrm{TV}(\bar{u}^n)$$

or

$$\mathrm{TV}(\bar{u}^{n+1}) \leqslant \mathrm{TV}(\bar{u}^{n}) + C\Delta t.$$

Both have the consequence that

$$\mathrm{TV}(\bar{u}^{n}) \leqslant \mathcal{C}(T)$$

for $n\Delta t \leqslant T$. TVD/TVB is also an important theoretical property: The space of all TVB functions is precompact, which has important consequences for convergence results.

This leads us to using a modified minmod function (min-mod-mod? min-mod^2? :-) Replace

$$\mathrm{minmod}(\tilde{u}_j, \bar{u}_{j+1} - \bar{u}_j, \bar{u}_j - \bar{u}_{j-1})$$

by

$$\overline{\mathrm{minmod}}(\tilde{u}_j, \bar{u}_{j+1} - \bar{u}_j, \bar{u}_j - \bar{u}_{j-1})$$

with

$$\overline{\mathrm{minmod}}(a,b,c) := \begin{cases} a & |a| \leqslant M\Delta x^2 \\ m(a,b,c) & \text{otherwise.} \end{cases}$$

We get the following properties:

- The scheme *is* TVB:

$$\mathrm{TV}(\bar{u}^{n+1}) \leqslant \mathrm{TV}(\bar{u}^{n+1}) + CM\,\Delta x^2\,N \leqslant \mathrm{TV}(\bar{u}^{n}) + C\Delta t$$

  where N is the total number of cells.

- The scheme maintains its high-order accuracy in smooth regions including at local extrema.

$$\tilde{u}_j = u_x(x_j)\frac{\Delta x}{2} + O(\Delta x^2) = O(\Delta x^2)$$

  near smooth extrema. The choice of M represents a tradeoff between oscillation and accuracy. One analysis of DG was carried out using $M = \frac{2}{3}|u_{x,x}|$ at extrema.

Discussion of HW#3, Problem 2: Here's how to show the CEI in the semidiscrete case. Let $f(\uparrow, \downarrow)$ and $U''(u) \geqslant 0$, and

$$F(u) = \int^u U'(u)f'(u)\mathrm{d}u \overset{\text{Integration by parts}}{=} U'(u)f(u) - \int^u U''(u)f(u).$$

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} + \frac{1}{\Delta x}\Big[\hat{f}(u_j, u_{j+1})_x - \hat{f}(u_{j-1}, u_j)\Big] \;=\; 0$$

Then

$$\frac{\mathrm{d}U(u_j)}{\mathrm{d}t} + \frac{1}{\Delta x}U'(u_j)\Big[\hat{f}(u_j, u_{j+1}) - \hat{f}(u_{j-1}, u_j)\Big] \;=\; 0.$$

Define

$$\hat{F}_{j+1/2} = U'(u_j)\hat{f}(u_j, u_{j+1}) - \int^{u_j} U''(u)f(u)\mathrm{d}u.$$

Then

$$\frac{\mathrm{d}U(u_j)}{\mathrm{d}t} + \frac{1}{\Delta x}\Big[\hat{F}_{j+1/2} - \hat{F}_{j-1/2}\Big] + \underbrace{\frac{1}{\Delta x}}_{\text{“junky” :)}}\Theta_j = 0.$$

Then

$$\begin{aligned}
\Theta_j &= \int^{u_j} U''(u)f(u)\mathrm{d}u - U'(u_j)\hat{f}(u_{j-1}, u_j) + U'(u_{j-1})\hat{f}(u_{j-1}, u_j) - \int^{u_{j-1}} U''(u)f(u)\mathrm{d}u \\
&= \int_{u_{j-1}}^{u_j} U''(u)f(u)\mathrm{d}u - (U'(u_j) - U'(u_{j-1}))\hat{f}(u_{j-1}, u_j) \\
&= \int_{u_{j-1}}^{u_j} U''(u)f(u)\mathrm{d}u - \int_{u_{j-1}}^{u_j} U''(u)\mathrm{d}u\hat{f}(u_{j-1}, u_j) \\
&= \int_{u_{j-1}}^{u_j} U''(u)\Big[f(u) - \hat{f}(u_{j-1}, u_j)\Big]\mathrm{d}u \geqslant 0.
\end{aligned}$$

Then

1. $u_{j-1} < u_j$. $u_{j-1} \leqslant u_j$

$$f(u) - \hat{f}(u_{j-1}, u_j) = \hat{f}(u, u) - \hat{f}(u_{j-1}, u_j) \ldots$$

...and then he cleaned the blackboard.

(End of HW discussion)

2.3 Essentially Non-Oscillatory Schemes

This scheme goes back to the idea of the MUSCL scheme,

$$u_{j+1/2}^{-} \leftarrow \bar{u}_j + \frac{1}{2}\mathrm{minmod}(\underbrace{\bar{u}_{j+1} - \bar{u}_j}_{\Delta_+}, \underbrace{\bar{u}_j - \bar{u}_{j-1}}_{\Delta_-}).$$

Recap: Newton interpolation. Suppose we have n points x_j with values y_j. Look for polynomial of degree $n-1$ such that $p(x_j) = y_j$. First review Lagrange polynomials and Lagrange interpolation $(l_i(x_j = \delta_{i,j})$. (omitted) Next up, Newton interpolation:

$$\begin{aligned} y[x_i] &= y_i \\ y[x_i, x_{i+1}] &= \frac{y[x_{i+1}] - y[x_i]}{x_{i+1} - x_i} \\ y[x_i, x_{i+1}, x_{i+2}] &= \frac{y[x_{i+1}, x_{i+2}] - y[x_i, x_{i+1}]}{x_{i+2} - x_i} \\ &\ldots \end{aligned}$$

Then

$$p(x) = y[x_0] + y[x_0, x_1](x - x_0) + y[x_0, x_1, x_2](x - x_0)(x - x_1) + y[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2).$$

But we are doing *reconstruction*, not interpolation. How can we convert reconstruction to interpolation? Consider that we're looking for a $p(x)$ such that

$$\frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} p(x) = \bar{u}_j \quad \text{for } j = 1, 2, \ldots m.$$

Then define

$$P(x) = \int_{x_{1/2}}^{x} p(\xi)\mathrm{d}\xi$$

and observe

$$P(x_{j+1/2}) = \int_{x_{1/2}}^{x_j+1/2} p(\xi)\mathrm{d}\xi = \sum_{l=1}^{j} \int_{x_{l-1/2}}^{x_{l+1/2}} \Delta x_l \bar{u}_l \quad j = 0, \ldots, m.$$

So how do we implement this? (Aargh, Fortran.) This algorithm works only for a uniform mesh:

1. Given the cell averages $\bar{u}_0$, $\bar{u}_1$, $\bar{u}_2$, ... as `ub(0),ub(1),...`
2. Compute the un-divided differences of $\bar{\boldsymbol{u}}$.

```
do i=1,n
u(i,0)=ub(1)
enddo
do l=1,m
do i =1,n-l
u(i,l)=u(i+1,l-1)-u(i,l-1)
enddo
enddo
```

3. At each location $j + 1/2$, to compute $u_{j+1/2}^{-}$, do

   a. Find the origin `is(j)` of the ENO stencil

```
is(j)=j
```

```
do l=1,m
if (abs( u(is(j)-1,l) ) .lt.  abs( u(is(j),l) ) ) is(j) = is(j)-1
enddo
```

b.

$$\underbrace{\texttt{un(j)}}_{u^-_{j+1/2}} = \sum_{\texttt{l=is(j)}}^{\texttt{is(j)+m}} \texttt{c(l-is(j),j-is(j-1))ub(l)}$$

(consider that $\texttt{l-is(j),j-is(j)} \in \{0, \ldots, \texttt{m}\}$).

2.4 Weighted ENO Schemes

Aside: Why is an interpolation polynomial monotone in the cell containing the discontinuity of a jump function? Suppose we're using 6 points, with the discontinuity in the middle cell. Then the polynomial is of degree five. The mean value theorem tells us that the derivative has zeros in the cells away from the discontinuity, of which there are four. But the derivative is of degree four, so it can at most have four zeros: Nice! There isn't one in the middle cell! (End aside)

Idea: Don't *choose* stencils like ENO, use a weighted sum.

Do it like this:

$$u^-_{j+1/2} = w_1 u^{(1)}_{j+1/2} + w_2 u^{(2)}_{j+1/2} + w_3 u^{(3)}_{j+1/2} + \cdots,$$

where $w_1 + w_2 + w_3 + \cdots = 1$ and $u^{(i)}_{j+1/2}$ are the higher-order linear reconstructions above. The goal is to choose the weights such that a higher order than just with $u^{(i)}_{j+1/2}$ is achieved, if the desired smoothness is available. Choose α_i such that the linear combination of smaller stencils adds up to a high-order stencil.

- $w_i = \alpha_i + O(\Delta x^2)$ in smooth regions
- If the stencil S_i contains a discontinuity, then we would like to have $w_i = O(\Delta x^4)$.

We define a "smoothness indicator", β_i to measure the smoothness of the function in stencil s_i.

$$\begin{aligned} \tilde{w}_i &= \frac{\alpha_i}{(\varepsilon + \beta_i)^2} \quad i = 1, 2, 3..., \quad \varepsilon = 10^{-6}, \\ w_i &= \frac{\tilde{w}_i}{\tilde{w}_1 + \tilde{w}_2 + \tilde{w}_3}. \end{aligned}$$

Shu's graduate student Jiang derived these smoothness indicators:

$$\beta_i = \Delta x^2 \int_{I_j} [(P'(x)^2 + \Delta x^2 (P''(x)^2)^2] \mathrm{d}x.$$

Homework:

- Code for Burgers':

$$\begin{cases} u_t + \left(\frac{u}{2}\right)^2_x = 0 \\ u(x,0) = 1 + \frac{1}{2}\sin(x) \end{cases}$$

  Give same output as before

  - ○ 3rd order linear using u

$$\begin{aligned} u^-_{j+1/2} &: \bar{u}_{j-1}, \bar{u}_j, \bar{u}_{j+1}, \\ u^+_{j+1/2} &: \bar{u}_j, \bar{u}_{j+1}, \bar{u}_{j+2}. \end{aligned}$$

  - ○ 3rd order TVD
  - ○ 3rd order TVB ($M = 5$)
  - ○ 3rd order ENO
  - ○ 5th order ENO
  - ○ 5th order WENO

Use 3rd order Runge-Kutta. (Might need to reduce Δt to see the 5th order accuracy.)
(Remember to initialize with and compare to cell averages of IC and exact solution!)

2.5 Finite Difference Methods

We are still considering

$$u_t + f(u)_x = 0,$$

which we hope to approximate by

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} + \frac{1}{\Delta x}\Big(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}\Big) = 0$$

using

$$\hat{f}_{j+1/2} = \hat{f}\,(u_{j-p}, \ldots, u_{j+q}).$$

Our requirements are

2.5.1 Accuracy

Accuracy means

$$\Big(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}\Big) = f(u)_x|_{x=x_j} + O(\Delta x^r).$$

Lemma 19. *(ENO paper by Shu, Osher) If there is a function $h(x)$ (which depends on Δx) s.t.*

$$f(u(x)) = \frac{1}{\Delta x}\int_{x-\Delta x/2}^{x+\Delta x/2} h(\xi)\mathrm{d}\xi,$$

then

$$f(u)_x = \frac{1}{\Delta x}\Big[h\Big(x + \frac{\Delta x}{2}\Big) - h\Big(x - \frac{\Delta x}{2}\Big)\Big].$$

All that's needed to obtain a higher-order scheme is now to approximate the function h to a certain degree of accuracy.

$$\{u_j\}\text{ given} \Rightarrow \{f(u_j)\}\text{ given} \Rightarrow \{\bar{h}_j\}\text{ given} \quad \underset{\text{reconstruction}}{\overset{\text{we want}}{\Longrightarrow}} \quad \{h_{j+1/2}\},$$

Then

$$f(u_j) = f(u(x_j)) = \frac{1}{\Delta x}\int_{x_j-\Delta x/2}^{x_j+\Delta x/2} h(\xi)\mathrm{d}\xi = \bar{h}_j.$$

2.5.2 Stability

For the moment, assume $f'(u) \geqslant 0$.

1. TVD Schemes:

    a. Use an upwind-biased stencil to compute $\hat{f}_{j+1/2}$, e.g.

    $$\{f(u_{j-1}), f(u_j), f(u_{j+1})\} \to \hat{f}_{j+1/2}.$$

    b. limit $\hat{f}_{j+1/2} - f(u_j) = \mathrm{d}f_j^+$.

    $$\mathrm{d}f_j^{+(\mathrm{mod})} = \mathrm{minmod}(\mathrm{d}f_j^+, f(u_{j+1}) - f(u_j), f(u_j) - f(u_{j-1})).$$

    Then

    $$\hat{f}_{j+1/2}^{(\mathrm{mod})} = f(u_j) + \mathrm{d}f_j^{+(\mathrm{mod})}.$$

Then use Harten's Lemma to prove TVD'ness. We only have the term $D_{j-1/2}$ since we have a unique wind direction by assumption, in

$$u_j^{n+1} = u_j^n - \lambda(-C_{j+1/2}(u_{j+1}^n - u_j^n) + D_{j-1/2}(u_j^n - u_{j-1}^n)).$$

By brute force, we have

$$\begin{aligned}
D_{j-1/2} &= \frac{f(u_j) + \mathrm{d}f_j^{+(\mathrm{mod})} - f(u_{j-1}) - \mathrm{d}f_{j-1}^{+(\mathrm{mod})}}{u_j - u_{j-1}}\\
&= \frac{f(u_j) - f(u_{j-1}) + \mathrm{d}f_j^{+(\mathrm{mod})} - \mathrm{d}f_{j-1}^{+(\mathrm{mod})}}{u_j - u_{j-1}}\\
&= \underbrace{\frac{f(u_j) - f(u_{j-1})}{u_j - u_{j-1}}}_{f'(\xi)} \left[1 + \underbrace{\frac{\mathrm{d}f_j^{+(\mathrm{mod})}}{f(u_j) - f(u_{j-1})}}_{0 \leqslant * \leqslant 1} - \underbrace{\frac{\mathrm{d}f_{j-1}^{+(\mathrm{mod})}}{f(u_j) - f(u_{j-1})}}_{0 \leqslant * \leqslant 1} \right]
\end{aligned}$$

with

$$0 \leqslant D_{j-1/2} \leqslant 2 \max_u |f'(u)|.$$

In order to lift the condition on the wind direction ($f'(u) \geqslant 0$), we need to consider only a *subclass* of montone fluxes, namely those characterized by *flux splitting*:

$$\hat{f}(u^-, u^+) = f^+(u^-) + f^-(u^+),$$

where

- $f(u) = f^+(u) + f^-(u)$
- $\dfrac{\mathrm{d}f^+(u)}{\mathrm{d}u} \geqslant 0, \quad \dfrac{\mathrm{d}f^-(u)}{\mathrm{d}u} \leqslant 0.$

One such example is Lax-Friedrichs: $f^{\pm}(u) = \frac{1}{2}(f(u) \pm \alpha u)$, where $\alpha = \max_u |f'(u)|$.

- Then use the previous (single-wind-direction) procedure w/ $f^+(u)$ instead of $f(u)$.
- The mirror-symetric (w.r.t. $j + 1/2$) procedure with $f^-(u)$ instead of $f^+(u)$.
- Thus we obtain $\hat{f}_{j+1/2}$.

Summary of FV versus FD:

FV	FD
$\bar{u}_j = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x,t)$	$u_j = u(x_j, t)$
reconstruction $\{\bar{u}_j\} \rightarrow \{u_{j\pm1/2}\}$	reconstruction $\{f^{\pm}(u_j)\} \rightarrow \{\hat{f}_{j+1/2}^{\pm}\}$
numerical flux $\hat{f}(u_{j+1}^-, u_{j+1}^+)$	numerical flux $\hat{f}_{j+1/2} = \hat{f}_{j+1/2}^+ + \hat{f}_{j+1/2}$
any $\hat{f}(\uparrow, \downarrow)$	*splittable* monotone flux $\hat{f}(u^-, u^+) = f^+(u^-) + f^-(u^+)$
Δx arbitrary (meshing unrestricted)	Δx uniform or smoothly mappable to uniform
	not much physics in the derivation

3 Two Space Dimensions

Now consider

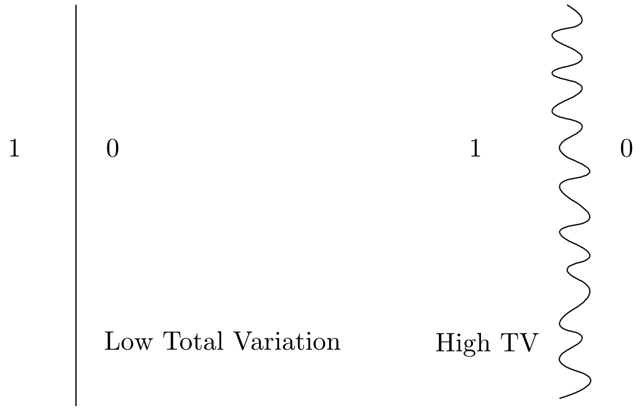
$$u_t + f(u)_x + g(u)_y = 0.$$

The good news are:

- Theoretical properties of weak solutions, entropy solutions etc. are the same as in 1D.
- All properties of monotone schmes (TVD, entropy condition, L_1-contraction, ...) are still valid in 2D.

Theorem 20. (Goodman & LeVeque) *In 2D, TVD schemes are at most first order accurate.*

Proof. (Very rough idea) Many things can happen in 2D:



□

“TVD” Schemes in the literature for nD means schemes which are TVD in 1D and are generalized to 2D in a dimension by dimension fashion, like this:

$$\begin{aligned}\frac{\mathrm{d}u_j}{\mathrm{d}t} + \frac{1}{\Delta x}\Big(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}\Big) &= 0\\ \text{with } \hat{f}_{j+1/2} \leftarrow \{f(u_{j-1}), f(u_j), f(u_{j+1})\}&\end{aligned}$$

becomes

$$\begin{aligned}\frac{\mathrm{d}u_{i,j}}{\mathrm{d}t} + \frac{1}{\Delta x}\Big(\hat{f}_{i+1/2,j} - \hat{f}_{i-1/2,j}\Big) + \frac{1}{\Delta y}\Big(\hat{f}_{i,j+1/2} - \hat{f}_{i,j-1/2}\Big) &= 0\\ \text{with } \hat{f}_{i+1/2,j} \leftarrow \{f(u_{i-1,j}), f(u_{i,j}), f(u_{i+1,j})\}.&\end{aligned}$$

They really are *not* TVD in more than one dimension.

One good property we have in more than one dimension is a *maximum principle*: Given a scheme in Harten form, i.e.

$$\begin{aligned}u_{i,j}^{n+1} &= u_{i,j}^n - \lambda_x\big[-C_{i+1/2,j}(u_{i+1,j}^n - u_{i,j}^n) + D_{i-1/2,j}(u_{i,j}^n - u_{i-1,j})\big]\\ &\qquad\;\; - \lambda_y\big[-C_{i,j+1/2}(u_{i,j+1}^n - u_{i,j}^n) + D_{i,j-1/2}(u_{i,j}^n - u_{i,j-1})\big]\end{aligned}$$

with

$$\begin{aligned}C_{i+1/2,j}, D_{i-1/2,j}, 1 - \lambda_x[C_{i+1/2,j} + D_{i+1/2,j}] &\geqslant 0,\\ C_{i,j+1/2}, D_{i,j-1/2}, 1 - \lambda_y[C_{i,j+1/2} + D_{i,j+1/2}] &\geqslant 0,\end{aligned}$$

we can proceed as follows:

$$\begin{aligned}u_{i,j}^{n+1} = &\underbrace{\big[1 - \lambda_x C_{i+1/2,j} - \lambda_x D_{i-1/2,j} - \lambda_y C_{i,j+1/2} - \lambda_y D_{i,j-1/2}\big]}_{\geqslant 0} u_{i,j}^n\\ &+ \lambda_x \underbrace{C_{i+1/2,j}}_{\geqslant 0} u_{i+1,j}^n + \lambda_x \underbrace{D_{i-1/2,j}}_{\geqslant 0} u_{i-1,j}^n\\ &+ \lambda_y \underbrace{C_{i,j+1/2}}_{\geqslant 0} u_{i,j+1}^n + \lambda_y \underbrace{D_{i,j-1/2}}_{\geqslant 0} u_{i,j-1}^n.\end{aligned}$$

Thus

$$\min(\text{stencil}) \leqslant u_{i,j}^{n+1} \leqslant \max(\text{stencil})$$

because it is a convex combination of the values in the stencil.

3.1 FV methods in 2D

Next, let’s consider FV methods in 2D. Let

$$\tilde{\bar{u}}_{i,j} = \frac{1}{\Delta x \Delta y}\int_{y_{j-1/2}}^{y_{j+1/2}}\int_{x_{i-1/2}}^{x_{i+1/2}} u(x, y, t)\mathrm{d}x\,\mathrm{d}y,$$

where we note that

$$\begin{aligned}\tilde{\cdot}\quad &\text{is}\quad \text{the cell-averaging operator in } y,\\ \bar{\cdot}\quad &\text{is}\quad \text{the cell-averaging operator in } x.\end{aligned}$$

Next,

$$\begin{aligned}&\frac{1}{\Delta x\Delta y}\int_{y_{j-1/2}}^{y_{j+1/2}}\int_{x_{i-1/2}}^{x_{i+1/2}} f(u)_x \mathrm{d}x\,\mathrm{d}y\\ =\;&\frac{1}{\Delta x\Delta y}\int_{y_{j-1/2}}^{y_{j+1/2}} f\left(u(x_{i+1/2},y,t)\right)-f(u(x_{i-1/2},y,t))\mathrm{d}y.\end{aligned}$$

Thus

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\bar{u}}_{i,j}+\frac{1}{\Delta x}\left[\frac{1}{\Delta y}\int_{y_{j-1/2}}^{y_{j+1/2}} f\left(u(x_{i+1/2},y,t)\right)\mathrm{d}y-\frac{1}{\Delta y}\int_{y_{j-1/2}}^{y_{j+1/2}} f\left(u(x_{i-1/2},y,t)\right)\mathrm{d}y\right]&\\ +\frac{1}{\Delta y}\left[\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} f\left(u(x,y_{j+1/2},t)\right)\mathrm{d}x-\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} f\left(u(x,y_{j-1/2},t)\right)\mathrm{d}x\right]&=0.\end{aligned}$$

The equality $(*)$ below is what breaks when we switch to a nonlinear equation.

FV Scheme:

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\bar{u}}_{i,j}+\frac{1}{\Delta x}\left[\hat{f}_{i+1/2,j}-\hat{f}_{i-1/2,j}\right]+\frac{1}{\Delta y}\left[\hat{g}_{i,j+1/2}-\hat{g}_{i,j-1/2}\right]=0.$$

3.1.1 The Linear Case

Let's consider a simple case to start:

$$u_t+a\,u_x+b\,u_y=0\quad\Rightarrow\quad f(u)=a\,u,\quad g(u)=b\,u.$$

In this case, we only have to perform *2 reconstructions per point*, so that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\bar{u}}_{i,j}+\frac{1}{\Delta x}\left[\underbrace{\frac{1}{\Delta y}\int_{y_{j-1/2}}^{y_{j+1/2}} f\left(u(x_{i+1/2},y,t)\right)\mathrm{d}y}_{\tilde{f}_{i+1/2,j}=a\tilde{u}_{i+1/2,j}\overset{(*)}{=}f(\tilde{u}_{i+1/2,j})}-\frac{1}{\Delta y}\int_{y_{j-1/2}}^{y_{j+1/2}} f\left(u(x_{i-1/2},y,t)\right)\mathrm{d}y\right]&\\ +\frac{1}{\Delta y}\left[\underbrace{\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} f\left(u(x,y_{j+1/2},t)\right)\mathrm{d}x}_{\bar{g}_{i,j+1/2}}-\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} f\left(u(x,y_{j-1/2},t)\right)\mathrm{d}x\right]&=0.\end{aligned}$$

3.1.2 The Nonlinear Case

In general, if $f(u)$ and $g(u)$ are nonlinear, then we have to perform one reconstructions for each point of the stencil, i.e. many times along one cut line through the stencil.

$$\begin{tikzcd}
 & \{\tilde{u}_{i+1/2,j}\} \arrow[r, "\text{1D rec}"] & \{u_{i+1/2,j+w_k}\} \arrow[r] & \{f(u_{i+1/2,j+w_k})\} \arrow[r, "\text{num.int.}"] & \{\hat{f}_{i+1/2,j}\} \\
\{\tilde{\bar{u}}_{i,j}\} \arrow[ur, "\text{1D rec}"] \arrow[dr, "\text{1D rec}"] & & & & \\
 & \{\bar{u}_{i,j+1/2}\} \arrow[r, "\text{1D rec}"] & \{u_{i+\omega_k,j+1/2}\} \arrow[r] & \{f(u_{i+\omega_k,j+1/2})\} \arrow[r, "\text{num.int.}"] & \{\hat{f}_{i,j+1/2}\}
\end{tikzcd}$$

Remark 21. These considerations only matter if we are interested in order of accuracy three or greater. If we are concerned with only second order accuracy, then

$$\tilde{\bar{u}}_{i,j}=u(x_i,y_j)+O(\Delta x^2,\Delta y^2)$$

is all we need.

3.2 Finite Difference Methods

We are still considering

$$u_t + f(u)_x + g(u)_y = 0,$$

but we switch the focus of our approximation to *actual point values*:

$$u_{i,j} = u(x_i, y_j, t)$$

to get the discretized conservation law

$$\frac{\mathrm{d}u_{i,j}}{\mathrm{d}t} + \frac{1}{\Delta x}\Big[\hat{f}_{i+1/2,j} - \hat{f}_{i-1/2,j}\Big] + \frac{1}{\Delta y}\big[\hat{g}_{i,j+1/2} - \hat{g}_{i,j-1/2}\big].$$

We need

$$\frac{1}{\Delta x}\Big[\hat{f}_{i+1/2,j} - \hat{f}_{i-1/2,j}\Big] = f(u)_x|_{x=x_j, y=y_j} + O(\Delta x^r, \Delta y^r)$$

for accuracy. This is identical to the 1D routine with fixed j.

4 Systems of Conservation Laws

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = \boldsymbol{0}$$

$\boldsymbol{u}$ is a vector, and so is $\boldsymbol{f}$. For the moment, x is still only 1-dimensional.

Example 22. Compressible flow:

$$\boldsymbol{u} = \left(\begin{array}{c}\rho\\ \rho v\\ E\end{array}\right), \quad \boldsymbol{f}(\boldsymbol{u}) = \left(\begin{array}{c}\rho v^2\\ \rho v + p\\ v(E+p)\end{array}\right),$$

where ρ is density, v is velocity, E is total energy and p is pressure. For a γ*-law gas*, for example, we could have the constitutive relationship

$$E = \frac{p}{\gamma - 1} + \frac{1}{2}\rho v^2.$$

E.g. for air $\gamma = 14$.

(Now, drop the bold-for-vector notation.)

4.1 A First Attempt: Generalize Methods from AM255

Example 23. (From 255) If $f(u) = A\,u$, then we have the equation

$$u_t + A\,u_x = 0 \tag{6}$$

If A has only real eigenvalues and a complete set of eigenvectors, then (6) is called *hypberbolic*. Consider

$$A\,r_i = \lambda_i r_i,$$

so that

$$A\,R = R\,\underbrace{\mathrm{diag}(\lambda_1, \ldots, \lambda_n)}_{\Lambda},$$

where R has the vectors r_i in its columns. Then we obtain

$$R^{-1} A\,R = \Lambda.$$

The rows l_i of R^{-1} are called the *left eigenvectors* of A, with $l_i A = \lambda_i l_i$ with $l_i r_j = \delta_{i,j}$.

Now, perform a change of variables, namely $v = R^{-1} u$, so that

$$v_t + \Lambda v_x = 0. \tag{7}$$

The goal for the nonlinear case is to take the lessons from the linear case, but rewrite the scheme (7) so that it only acts on u. If all the *eigenvalues are positive*, then we can rewrite the upwind scheme (now reinstating bold-face-for-vector, with index for x location)

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{v}_j}{\mathrm{d}t} + \frac{1}{\Delta x}\Lambda[\boldsymbol{v}_j - \boldsymbol{v}_{j-1}] &= \mathbf{0} \\ \Leftrightarrow \frac{\mathrm{d}\boldsymbol{u}_j}{\mathrm{d}t} + \frac{1}{\Delta x}R\Lambda R^{-1}[\underbrace{R\boldsymbol{v}_j}_{\boldsymbol{u}_j} - R\boldsymbol{v}_{j-1}] &= \mathbf{0} \\ \Leftrightarrow \frac{\mathrm{d}\boldsymbol{u}_j}{\mathrm{d}t} + \frac{1}{\Delta x}A[\boldsymbol{u}_j - \boldsymbol{u}_{j-1}] &= \mathbf{0}. \end{aligned}$$

If we do not have the above eigenvalue condition, then we need a good way to write the resulting system concisely. Why not start with some notation...

$$a^+ := \begin{cases} a & a \geqslant 0, \\ 0 & \text{otherwise}, \end{cases} \qquad a^- := \begin{cases} 0 & \text{otherwise} \\ a & a \leqslant 0 \end{cases}.$$

Thus $|a| = a^+ - a^-$ and $a = a^+ + a^-$. This notation has natural generalizations to matrices and vectors. We obtain the following scheme in $\boldsymbol{v}$:

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{v}_j}{\mathrm{d}t} + \frac{1}{\Delta x}\{\Lambda^+[\boldsymbol{v}_j - \boldsymbol{v}_{j-1}] + \Lambda^-[\boldsymbol{v}_{j+1} - \boldsymbol{v}_j]\} &= \mathbf{0} \\ \Leftrightarrow \frac{\mathrm{d}\boldsymbol{u}_j}{\mathrm{d}t} + \frac{1}{\Delta x}\left\{\underbrace{R\Lambda^+R^{-1}}_{A^+:=}[\boldsymbol{u}_j - \boldsymbol{u}_{j-1}] + \underbrace{R\Lambda^-R^{-1}}_{A^-:=}[\boldsymbol{u}_{j+1} - \boldsymbol{u}_j]\right\} &= \mathbf{0}. \end{aligned}$$

Note the slightly ambiguous notation here–A^+ is not the positive part of A in the above sense, even though $A = A^+ + A^-$ still holds.

4.2 How to Generalize Scalar Higher-Order Schemes to Systems

We are still considering

$$u_t + A\,u_x = 0.$$

1. Find the eigenvalues of A, hence Λ
   Also find the eigenvectors of A, hence R and R^{-1}.
2. At each point that we need to compute a flux or a reconstruction, say at $x_{j+1/2}$, do the following
   a. $\boldsymbol{v}_i = R^{-1}\boldsymbol{u}_i$ $(i = j - p, \ldots j + q)$
   b. Use the scalar subroutine to each component of $\boldsymbol{v}$ to obtain a reconstruction $\boldsymbol{v}_{j+1/2}$.
   c. $\boldsymbol{u}_{j+1/2} = R\boldsymbol{v}_{j+1/2}$.

Now, why should we do this transformation instead of just applying the scalar subroutine to $\boldsymbol{u}$? Consider this example:

$$\begin{aligned} (v_1)_t + (v_1)_x &= 0, \\ (v_2)_t + (v_2)_x &= 0. \end{aligned}$$

Any combination of u is bound to develop *two shocks*, travelling at different speeds. If however we calculate v, then we retain the two nicely separated shocks. To drive home the point, ENO always counts on the fact that it can find a stencil near a shock where the function is smooth. For a point "*trapped*" between two shocks, this assumption is violated, and we will lose something.

Also note that this procedure only makes sense if you are doing something nonlinear in step 2b.

Next, note that if our discussion is targetted at generalizing to nonlinear conservation laws. Consequently, it is really pointless to actually carry out steps 2a and 2c each time unless the matrix A is actually changing as it will be.

Note 24. "Theorem": All results about stability and convergence carry over to the case of *linear systems* if the numerical schemes use the above the "characteristic" procedure.

4.3 The Nonlinear Case

If we consider the equation

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = \boldsymbol{0},$$

then

- There is essentially no theory.
- The numerical procedure is essentially identical to that for the linear system case performed in (local) characteristic fields.

Additional Homework: (This+HW4 due Nov 29)

1. Add third order finite difference version to HW4.

[one class's worth of material is missing here. It is available as a separate PDF file called `257-missed-class.pdf` courtesy of Ishani Roy.]

5 The Discontinuous Galerkin Method

$$u_t + f(u)_x = 0.$$

To begin a FV discretization, we rewrite this as

$$\frac{1}{\Delta t}\int_{x_{j-1/2}}^{x_{j+1/2}} (u_t + f(u)_x)\mathrm{d}x = 0,$$

which results in:

$$\frac{\mathrm{d}\bar{u}_j}{\mathrm{d}t} + \frac{1}{\Delta x_j}\big(f(u_{j+1/2}) - f(u_{j-1/2})\big) = 0$$

FV in its full glory is

$$\frac{\mathrm{d}\bar{u}_j}{\mathrm{d}t} + \frac{1}{\Delta x_j}\Big(\hat{f}(u_{j+1/2}^-, u_{j+1/2}^+) - \hat{f}(u_{j-1/2}^-, u_{j-1/2}^+)\Big),$$

where, to make this a scheme, we need a monotone flux $\hat{f}(u^-, u^+)$, which needs to satisfy the following criteria:

- $\hat{f}(\uparrow, \downarrow)$,
- $\hat{f}(u, u) = u$,
- Lipschitz continuous.

For DG, we do something different. We multiply the PDE by a "test function" v, then integrate the result over the interval $(x_{j-1/2}, x_{j+1/2})$

$$\int_{x_{j-1/2}}^{x_{j+1/2}} (u_t + f(u)_x)v\,\mathrm{d}x = 0.$$

Now consider u and v both from a finite-dimensional function space V_h, where $h = \max\,(x_{j+1/2}, x_{j-1/2})$. The space is then given by

$$V_h = \{w \colon w|_{I_j} \in \mathcal{P}^k(I_j)\},$$

where $I_j = (x_{j-1/2}, x_{j+1/2})$ and $\mathcal{P}^k(I_j)$ is a collection of polynomials of degree $\leqslant k$ on cell I_j. We observe $\dim V_h = N\cdot(k+1)$. Then perform integration by parts and write

$$\int_{x_{j-1/2}}^{x_{j+1/2}} u_t v - \int_{x_{j-1/2}}^{x_{j+1/2}} f(u)v_x\,\mathrm{d}x + f(u_{j+1/2})v_{j+1/2} - f(u_{j-1/2})v_{j-1/2} = 0.$$

To make this into a scheme: find $u \in V_h$ such that

$$\int_{I_j} u_t v\,\mathrm{d}x - \int_{I_j} f(u)v_x\mathrm{d}x + \underbrace{f(u_{j+1/2})v_{j+1/2} - f(u_{j-1/2})v_{j-1/2}}_{?} = 0$$

is true for any test function $v \in V_h$. But the term marked "?" is meaningless, since the functions are double-valued at the spots in question. To motivate a meaning for the term, consider the following: If we take the test function

$$v = \begin{cases} 1 & x \in I_j, \\ 0 & \text{elsewhere}, \end{cases}$$

we recover

$$\begin{aligned} \int_{I_j} u_t\,\mathrm{d}x + f(u_{j+1/2})\underbrace{v_{j+1/2}}_{\text{from left}} - f(u_{j-1/2})\underbrace{v_{j-1/2}}_{\text{from right}} &= 0 \\ \int_{I_j} u_t\,\mathrm{d}x + f(u_{j+1/2}) - f(u_{j-1/2}) &= 0, \end{aligned}$$

which is exactly reminiscent of the FV scheme, motivating the equality

$$f(u_{j+1/2}) - f(u_{j-1/2}) = \hat{f}(u_{j+1/2}^-, u_{j+1/2}^+) - \hat{f}(u_{j-1/2}^-, u_{j-1/2}^+)$$

and thus the scheme

$$\int_{I_j} u_t v\,\mathrm{d}x - \int_{I_j} f(u) v_x \mathrm{d}x + \hat{f}(u_{j+1/2}^-, u_{j+1/2}^+) v_{j+1/2}^- - \hat{f}(u_{j-1/2}^-, u_{j-1/2}^+) v_{j-1/2}^+ = 0.$$

Pick a basis for V_h:

$$V_h = \{\varphi_j^{(l)}: 1 \leqslant j \leqslant N, 0 \leqslant l \leqslant k\}.$$

For example, we could take

$$\begin{array}{rcl} \varphi_j^{(0)}(x) & = & \mathbf{1}_{I_j}(x), \\ \varphi_j^{(1)}(x) & = & (x - x_j)\mathbf{1}_{I_j}(x), \\ \varphi_j^{(2)}(x) & = & (x - x_j)^2 \mathbf{1}_{I_j}(x), \\ & \vdots & \end{array}$$

then

$$u(x,t) = \sum_{l=1}^{k} u_j^{(l)}(t) \varphi_j^{(l)}(x), \quad x \in I_j.$$

Now take $v = \varphi_j^{(m)}(x)$, $m = 0, 1, \ldots, l$ and put that into our scheme

$$\begin{array}{rcl} & & \displaystyle\int_{x_{j-1/2}}^{x_{j+1/2}} \left(\sum_{l=0}^{k} u_j^{(l)}(t) \varphi_j^{(l)}(x) \right)_t \varphi_j^{(m)}(x) \mathrm{d}x \\ & & \displaystyle - \int_{x_{j-1/2}}^{x_{j+1/2}} f\left(\sum_{l=0}^{k} u_j^{(l)}(t) \varphi_j^{(l)}(x) \right) \frac{\mathrm{d}}{\mathrm{d}x} \varphi_j^{(m)}(x) \mathrm{d}x \\ & & \displaystyle + \hat{f}\left(\sum_{l=0}^{k} u_j^{(l)}(t) \varphi_j^{(l)}(x_{j+1/2}), \sum_{l=0}^{k} u_{j+1}^{(l)}(t) \varphi_{j+1}^{(l)}(x_{j+1/2}) \right) \varphi_j^{(m)}(x_{j+1/2}) \\ & & \displaystyle - \hat{f}\left(\sum_{l=0}^{k} u_{j-1}^{(l)}(t) \varphi_{j-1}^{(l)}(x_{j-1/2}), \sum_{l=0}^{k} u_j^{(l)}(t) \varphi_j^{(l)}(x_{j-1/2}) \right) \varphi_j^{(m)}(x_{j-1/2}) \; = \; 0. \end{array}$$

Working with that yields

$$\begin{array}{rcl} \displaystyle\sum_{l=0}^{k} \frac{\mathrm{d}}{\mathrm{d}t} u_j^{(l)}(t) \underbrace{\int_{x_{j-1/2}}^{x_{j+1/2}} \varphi_j^{(l)}(x) \varphi_j^{(m)}(x) \mathrm{d}x}_{(k+1)\times(k+1) \text{ matrix}} & & \\ + F(\boldsymbol{u}_{j-1}(t), \boldsymbol{u}_j(t), \boldsymbol{u}_{j+1}(t)) & = & 0, \end{array}$$

where

$$\boldsymbol{u}_j(t) = \begin{pmatrix} u_j^{(0)}(t) \\ \vdots \\ u_j^{(k)}(t) \end{pmatrix}.$$

If the matrix above (also called the *local mass matrix*) is, we can rewrite the scheme as

$$\sum_{l=0}^{k} \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{u}_j(t) + \tilde{\boldsymbol{F}}(\boldsymbol{u}_{j-1}(t), \boldsymbol{u}_j(t), \boldsymbol{u}_{j+1}(t)) = 0,$$

which, if $\tilde{\boldsymbol{F}}$ is locally Lipschitz (which it is), gives a well-defined scheme. If we have a linear PDE $f(u) = A\,u$, where $A = A(x,t)$, then the scheme becomes

$$\frac{\mathrm{d}\boldsymbol{u}_j(t)}{\mathrm{d}t} + [B_{j-1} \boldsymbol{u}_{j-1} + C_j \boldsymbol{u}_j(t) + D_{j+1} \boldsymbol{u}_{j+1}(t)] = 0,$$

where the three matrices B_{j-1}, C_j, D_{j+1} (each of size $(k+1) \times (k+1)$) do not depend on $\boldsymbol{u}$.

5.1 Some Theoretical Properties of the Scheme

This scheme satisfies the cell entropy inequality for the square entropy $U(u) = u^2/2$. Recall the general entropy inequality, where for an entropy U satisfying $U''(u) \geqslant 0$ and a matching flux

$$F(u) = \int^u U'(u) f'(u) \mathrm{d}u,$$

we have

$$U(u)_t + F(u)_x \leqslant 0$$

in some weak sense.

Proof. Take $v = u$ in the scheme:

$$\begin{aligned}
\int_{I_j} u_t u \,\mathrm{d}x - \int_{I_j} \underbrace{f(u) u_x}_{g(u)_x} \mathrm{d}x + \hat{f}_{j+1/2} u_{j+1/2}^- - \hat{f}_{j-1/2} u_{j-1/2}^+ &= 0\\
\frac{\mathrm{d}}{\mathrm{d}t}\left(\int \frac{u^2}{2} \mathrm{d}x \right) - g(u_{j+1/2}^-) + g(u_{j-1/2}^+) + \hat{f}_{j+1/2} u_{j+1/2}^- - \hat{f}_{j-1/2} u_{j-1/2}^+ &= 0\\
\frac{\mathrm{d}}{\mathrm{d}t}\left(\int \frac{u^2}{2} \mathrm{d}x \right) + \hat{F}_{j+1/2} - \hat{F}_{j-1/2} + \underbrace{\left[-g(u_{j-1/2}^-) + \hat{f}_{j-1/2} u_{j-1/2}^- + g(u_{j-1/2}^+) - \hat{f}_{j-1/2} u_{j-1/2}^+ \right]}_{\Theta_{j-1/2}} &= 0
\end{aligned}$$

where we have taken

$$g(u) = \int^u f(u) \mathrm{d}u, \quad g'(u) = f(u)$$

and

$$\hat{F}_{j+1/2} = -g(u_{j+1/2}^-) + \hat{f}_{j+1/2} u_{j+1/2}^-,$$

where we observe that $\hat{F}$ is consistent, i.e.

$$\begin{aligned}
\hat{F}(u,u) &= -g(u) + f(u) u\\
&\overset{?}{=} \int^u u\, f'(u)\, \mathrm{d}u\\
&\leftarrow \int^u u \mathrm{d} f(u) = u\, f(u) - \underbrace{\int^u f(u) \mathrm{d}u}_{g(u)}.\\
\hat{F}' &= -f(u) + f'(u) u + f(u) = f'(u) u.
\end{aligned}$$

We would like to show $\Theta_{j-1/2} \geqslant 0$ to prove the cell entropy inequality, i.e. the term above $\leqslant 0$.

$$\begin{aligned}
\Theta &= -g(u^-) + \hat{f}(u^-, u^+) u^- + g(u^+) - \hat{f}(u^-, u^+) u^+\\
&= g(u^+) - g(u^-) - \hat{f}(u^-, u^+)(u^+ - u^-)\\
&= g'(\xi)(u^+)(u^+ - u^-) - \hat{f}(u^-, u^+)(u^+ - u^-)\\
&= (u^+ - u^-)(f(\xi) - \hat{f}(u^-, u^+))\\
&= (u^+ - u^-)(\hat{f}(\xi, \xi) - \hat{f}(u^-, u^+)).
\end{aligned}$$

After a simple case distinction on $u^- \lesseqgtr \xi \lesseqgtr u^+$ and using $\hat{f}(\uparrow, \downarrow)$, we find $\Theta \geqslant 0$. □