High-order Finite Volume Methods

Oguz Ziya Koseomur, *Computational Science and Engineering, Technical University Munich*

Abstract

In this review, the main ingredients of high-order finite volume methods are reviewed. First, Godunov's method for the linear systems, the *Reconstruct-Evolve-Average (REA)* algorithm is introduced. This algorithm is the basis for the development of most high order methods and it is mainly based on the polynomial reconstruction of the data. Then, the monotonic upstream-centered scheme for conservation laws (MUSCL) as well as the total variation and the total variation diminishing (TVD) methods are discussed, which are aimed to bound the oscillations in the solution. A method to obtain higher resolution in the temporal domain, the semi-discrete form, is introduced. Semi-discrete form allows us to decouple issues of the temporal discretization and the spatial discretization. Finally, more advanced methods to reconstruct the higher order polynomials, such as Newton's divided differences, essentially non-oscillatory schemes are briefly mentioned. Most of the discussions and derivations are centered around the linear equations.

Index Terms

finite volumes, high-order, limiters, TVD, total variation, MUSCL, ENO, WENO

I. Introduction

THE upwind method for hyperbolic PDEs showed great success overcoming the instabilities introduced with central schemes. However, a trade-off for the accuracy has been made to make the solutions more stable. In addition, this method is highly dissipative, therefore it tends to smooth the solution, which is not accurate for the discontinuous solutions. These discontinuities can either be introduced as initial conditions or the artifacts of the non-linear equations. Considering this, we expect a numerical method to behave well in smooth regions as well as in discontinuous regions. Since the finite volume methods consists of several steps, it is not trivial to introduce the stable higher order methods, unlike finite differences. A special care must be taken in each step of the development of the solution, such as representing the data as a continuous set of points, while not smoothing the data in the discontinuous regions, accomplishing integration and differentiation operations in both spatial and temporal domains without loss of accuracy. In addition, we should also consider their effects on each other too. As a result, a systematic, reproducible procedure with a mathematical background is necessary to develop the high-order methods.

Most of the methods are tested and compared using the linear advection equation with constant velocity. The advection velocity, $\bar{u}$, is assumed to be positive and constant in order to reduce the complexity in the derivations.

$$q_t + \bar{u} q_x = 0 \tag{1}$$

In the following sections, first, the main algorithm for the most of the methods, *reconstruct-evolve-average* algorithm and a second order method and its drawbacks, the Lax-Wendroff method are introduced. Then, the methods to blend high order methods with the lower order methods without adding oscillations to our solutions, the limiters and MUSCL method, are discussed. Methods for the temporal high resolution are introduced and the paper is finalized with the more advanced methods for the polynomial reconstruction, ENO and WENO methods.

II. Reconstruct-Evolve-Average (REA) Algorithm

REA algorithm is the upwind method for systems of equations and plays an important role to develop high-order methods. The method consists of three main steps:

1) **Reconstruct:** On the discretized domain, construct a piecewise polynomial using the cell averages. Denoting P as the piecewise polynomial constructed from cell averages, Q_i^n,

$$\tilde{q}^n(x, t_n) = P(x)$$

2) **Evolve:** In this step, time marching for one time step is performed for the reconstructed state.

$$\tilde{q}^n(x, t_n) \longrightarrow \tilde{q}^n(x, t_{n+1})$$

3) **Average:** Reconstruction of the cell values requires the average values for each cell on cell centers. Therefore, the evolved state is averaged over each cell, in a conservative fashion.

$$Q_i^{n+1} = \frac{1}{\Delta x} \int_{C_i} \tilde{q}^n(x, t_{n+1}) dx$$

Then the algorithm repeats itself.

The simplest choice for the reconstruction in the step 1 is using a piecewise constant polynomial to represent the data, which is also the original version of the algorithm. This construction leads to the basic Riemann problems. However, this approach only results in first order accuracy. In order to obtain better accuracy, one should use better reconstruction approaches, which are discussed in the next sections. The finite volume realization of this algorithm with the numerical flux can be derived as follows in Equations 2 to 4. Recalling the numerical flux definition:

$$F^n_{i-1/2} \approx \frac{1}{\Delta x}\int_{t_n}^{t_{n+1}} f(q(x_{i-1},t))dt$$

which approximates the flux between the cells $i-1$ and i. Instead of using the exact value $q(x,t)$ in the integral, piecewise version $\tilde{q}(x,t)$ can be used, in which the calculation of the respective integral is exact. Then the algorithm reduces to:

1) Solve the Riemann problem on the interface $x_{i-1/2}$
$$Q^*_{i-1/2} = R(Q_{i-1}, Q_i) \tag{2}$$
where the R represents an exact or approximate Riemann solver and the $Q^*_{i-1/2}$ represents the solution of the Riemann problem.
2) Calculate the flux as
$$F_{i-1/2} = F(Q_{i-1}, Q_i) = f(Q^*_{i-1/2}) \tag{3}$$
3) Apply flux-differencing formula to obtain the next state
$$Q^{n+1}_i = Q^n_i - \frac{\Delta t}{\Delta x}(F_{i+1/2} - F_{i-1/2}) \tag{4}$$

Consider the linear advection equation given in the Eq. 1. With the first order upwind method, Equation 4 reduces to:

$$Q^{n+1}_i = Q^n_i - \frac{\bar{u}\Delta t}{\Delta x}(Q^n_i - Q^n_{i-1})$$

which is the update formula for the first-order upwind method of Godunov, and the most common presentation of the REA algorithm.

III. Lax-Wendroff Method

The Lax-Wendroff method is a second order method which is derived from the Taylor series expansion. For the linear advection equation in Eq. 1, differentiation of this equation with respect to t gives

$$q_{tt} = -\bar{u}q_{xt} = -\bar{u}(-\bar{u}q_x)_x = \bar{u}^2 q_{xx}$$

since the $q_{xt} = q_{tx}$. As for the next step, we can use the Taylor series expansion in the temporal domain to expand the next time step

$$q(x,t_{n+1}) = q(x,t_n) + \Delta t q_t(x,t_n) + \frac{1}{2}(\Delta t)^2 q_{tt}(x,t_n) + O(\Delta t^3)$$

Now, we can replace the derivatives with respect to t to x using the previous relation as

$$q(x,t_{n+1}) = q(x,t_n) - \bar{u}\Delta t q_x(x,t_n) + \frac{1}{2}(\Delta t)^2 \bar{u}^2 q_{xx}(x,t_n) + O(\Delta t^3)$$

Finally, keeping only up to the third order error term and replacing the differentials with the central differences and exact values with the cell averages, we obtain the update formula of the Lax-Wendroff method.

$$Q^{n+1}_i = Q^n_i - \frac{\bar{u}\Delta t}{2\Delta x}(Q^n_{i+1} - Q^n_{i-1}) + \frac{1}{2}\left(\frac{\bar{u}\Delta t}{\Delta x}\right)^2 (Q^n_{i+1} - 2Q^n_i + Q^n_{i-1})$$

The Lax-Wendroff method is a second order method as opposed to the first order upwind method. However, in case of a discontinuity in the solution, it begins to produce unphysical oscillations. This is mainly due to the dispersive nature of the error term, third derivative. In addition, this dispersive error also causes some kind of phase shift even in the smooth regions. For example, Figure 2 compares the first order-upwind method and the Lax-Wendroff method on advection equation. While the first order upwind method results in dissipative results, the Lax-Wendroff method shows highly oscillatory behaviour.

To summarize, we get a highly dissipative, low order but non-oscillatory solution with upwind method, while we get a high order but oscillatory solution near discontinuities. The idea of limiting tries to combine the best features of both methods: high order solution in the smooth regions, non-oscillatory solution near discontinuities. Keeping this in mind, the Lax-Wendroff method is an important method to characterize higher order methods developed later in this paper.

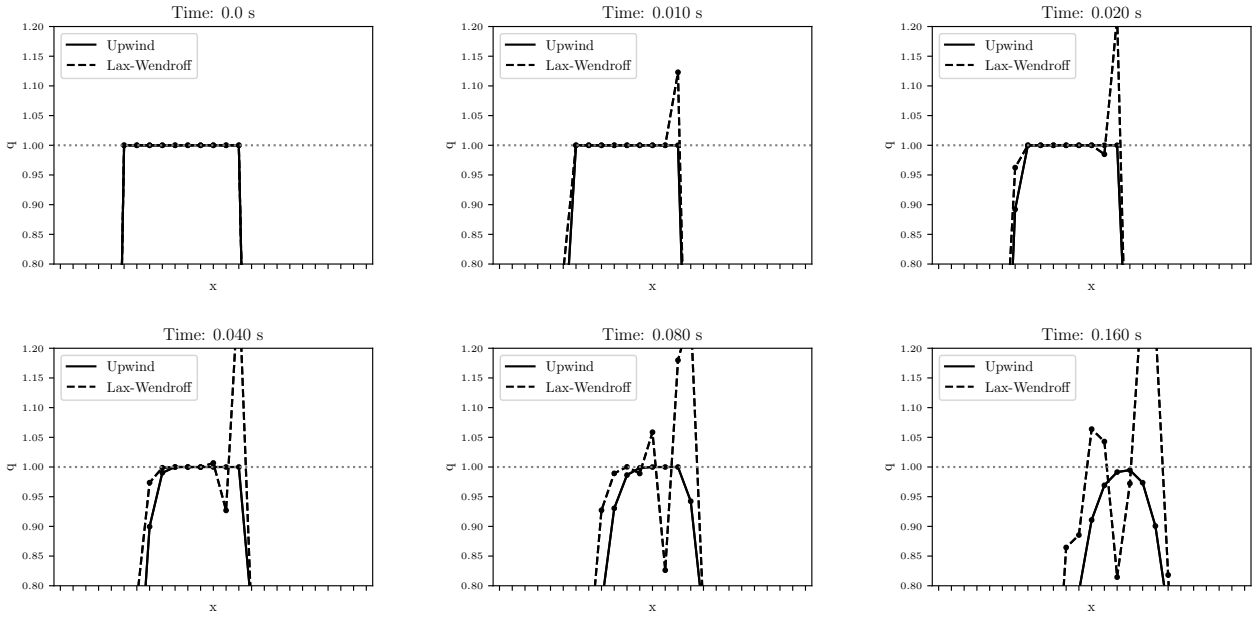

Fig. 1. Evolution of the step initial condition under the advection equation with the first order upwind and Lax-Wendroff methods. CFL number is 0.3. The highly oscillatory behaviour of the Lax-Wendroff method is clearly visible from the first time step.

IV. Reconstruction

In the REA algorithm, cell averages are reconstructed using the piecewise constant polynomial. However, it appears that this reconstruction is one of the limiting cases to obtain high order methods. Instead of the piecewise constant reconstruction, a better, and simple reconstruction is linear piecewise reconstruction, such that:

$$\tilde{q}^n(x, t_n) = Q_i^n + \sigma_i^n (x - x_i)$$

Two main properties of this reconstruction are the value at the cell center is equal to Q_i^n. In addition, the integral over the cell is equal to Q_n^i, no matter what the slope σ_i^n is and this is crucial to obtain conservative schemes.

After a straightforward but long derivation with the piecewise linear reconstruction, one can obtain the following general form:

$$Q_i^{n+1} = Q_i^n - \frac{\bar{u}\Delta t}{\Delta x}(Q_i^n - Q_{i-1}^n) - \frac{1}{2}\frac{\bar{u}\Delta t}{\Delta x}(\Delta x - \bar{u}\Delta t)(\sigma_i^n - \sigma_{i-1}^n) \tag{5}$$

It is simply the first order upwind method with correction which depends on the slopes.

A. How to Choose Slopes?

Currently, the slopes are the free parameters for the Equation 2. As in derivative approximation, one can choose the following slopes without loss of generality

Centered Slope (CS)	Upwind Slope (US)	Downwind Slope (DS)
$\sigma_i^n = (Q_{i+1}^n - Q_{i-1}^n)/(2\Delta x)$	$\sigma_i^n = (Q_i^n - Q_{i-1}^n)/(\Delta x)$	$\sigma_i^n = (Q_{i+1}^n - Q_i^n)/(\Delta x)$

The main downside of these slopes is that they are derived such that the solution is smooth. Despite their simplicity, this assumption is invalid near discontinuities and they tend to overshoot the values depending on their direction bias, which grow over the time and create oscillatory solutions. Therefore, we need to focus on the slopes which do not add additional oscillations.

B. Total Variation

In the previous section, the overshooting behaviour of the trivial slope choices are mentioned. In order to choose more sophisticated slopes which do not have oscillatory behaviour, a function that measures oscillations should be derived. This function is denoted as Total Variation (TV) and defined as:

$$TV(Q) = \sum_{-\infty}^{\infty} |Q_i - Q_{i-1}|$$

This function basically represents how oscillatory the solution is. For example, the linear advection equation with constant velocity cannot introduce additional oscillations in its nature, so the total variation of the linear advection equation is expected to be constant over time. Therefore, the numerical method used to solve this problem should not introduce additional oscillations too. Based on this, a numerical method is called total variation diminishing (TVD) if

$$TV(Q^n) \geq TV(Q^{n+1})$$

This property is extremely important for a method to have non-oscillatory characteristic.

C. MUSCL-Type Schemes

Monotone Upstream-Centered Scheme for Conservation Laws, MUSCL, type schemes are the the general category of the methods in which the degree of the reconstructed polynomial is higher than the zero, which means non-constant reconstruction. Another property of the MUSCL schemes is that the higher order reconstruction should not introduce additional oscillations, as mentioned in the trivial slope selections.

When the reconstruction is not constant, the definition of the Riemann problem is violated. Therefore, the Generalized Riemann Problem is defined such that the values at the cell centers are considered to compute the in-cell fluxes for the problem

$$q_t + f(q) = 0$$

$$q(x,0) = \begin{cases} q_i^R(x), & x < 0 \\ q_{i+1}^L(x), & x > 0 \end{cases}$$

Moreover, the MUSCL-type schemes are not only limited with the linear reconstruction, also the higher-order reconstructions are possible.

A specific type of the MUSCL-type method, the MUSCL-Hancock method follows the algorithmic procedure given in equations 6 to 10.

1) Reconstruct the data with the linear reconstruction:

$$Q_i^L = Q_i^n - \frac{1}{2}\Delta x \tag{6}$$

$$Q_i^R = Q_i^n + \frac{1}{2}\Delta x \tag{7}$$

2) Evolve the solution by a time $0.5\Delta t$

$$\bar{Q}_i^L = Q_i^L + \frac{1}{2}\frac{\Delta t}{\Delta x}(F_{i-1/2} - F_{i+1/2}) \tag{8}$$

$$\bar{Q}_i^R = Q_i^R + \frac{1}{2}\frac{\Delta t}{\Delta x}(F_{i-1/2} - F_{i+1/2}) \tag{9}$$

3) Solve the Generalized Riemann problem on the interface with the piecewise constant data

$$q(x,0) = \left\{ \begin{array}{ll} Q_i^R(x), & x < 0 \\ Q_{i+1}^L(x), & x > 0 \end{array} \right\} \tag{10}$$

D. TVD Limiters

Considering TVD property, more sophisticated slopes can be derived, such as

MinMod	SuperBee	MC
$minmod(US, DS)$	$maxmod[minmod(DS, 2US), minmod(2DS, US)]$	$minmod(CS, 2US, 2DS)$

The MinMod slope calculates the both upwind and downwind slopes and chooses the smaller one to introduce less oscillations. If the signs are different for these slopes, then it decides on the slope of zero. This results in sharper reconstruction of the discontinuity without introducing additional oscillations. Similarly, the SuperBee slope also computes different sided slopes, but it chooses the one with the larger modulus. Monotonized central-difference limiter (MC) is one of the default choices for the most of the problems, since it chooses the central scheme when the solution is smooth therefore it does not introduce artificial effects on the smooth regions.

Although the formulation for the slope-limiting is trivial and intuitive, it is not in the flux-formulation yet. Beginning from the slope-limiting, the general formulation for the flux-limited numerical methods can be derived. In this way, we can obtain the flux-limiter counterparts of the slope-limiters. The key point is that now, we associate the slope with the interface rather than the each cell. Defining the jump between the cells i and $i-1$:

$$\Delta Q_{i-1/2}^n = Q_i^n - Q_{i-1}^n$$

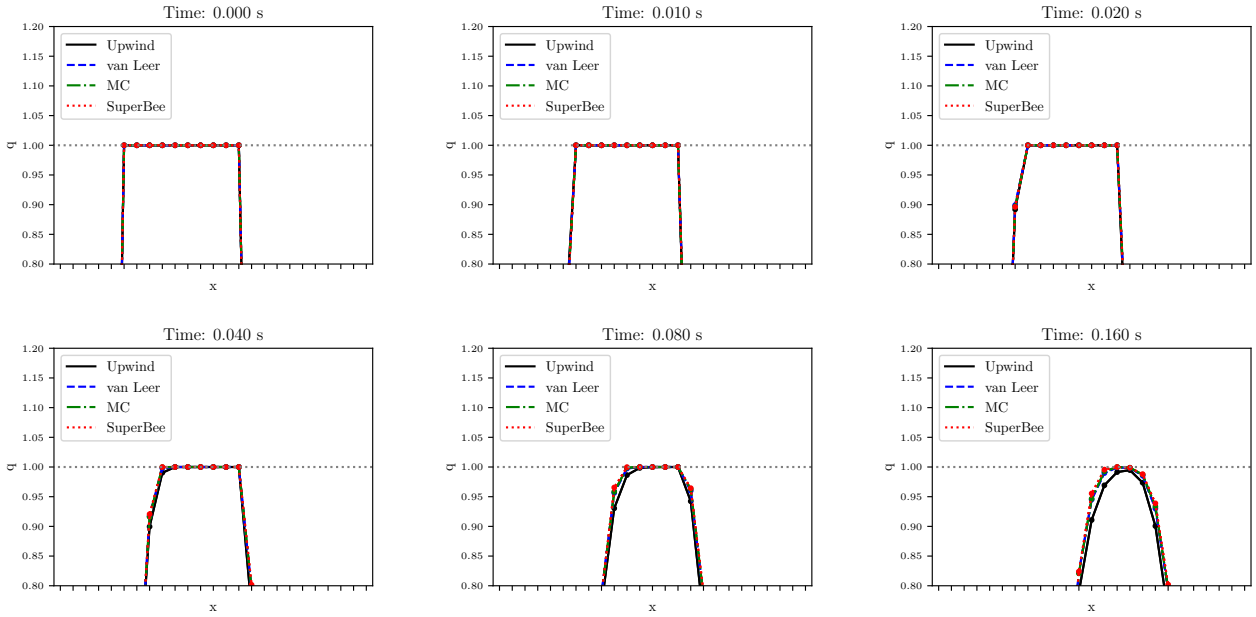

Fig. 2. Evolution of the step initial condition under the advection equation with the different limiters. CFL number is 0.3. The least dissipative limiters are the superbee and van Leer while the difference is not significant.

Then the limited version of the flux on this face can be written as (given that $\bar{u} > 0$):

$$F^n_{i-1/2} = \bar{u}Q^n_i + \frac{1}{2}\bar{u}\left(1 - \frac{u\bar{\Delta}t}{\Delta x}\right)\phi(\theta^n_{i-1/2})\Delta Q^n_{i-1/2}$$

where the ϕ is the limiter function while the θ determines the smoothness and is defined as:

$$\theta^n_{i-1/2} = \frac{\Delta Q^n_{i-3/2}}{\Delta Q^n_{i-1/2}}$$

The flux-limiters defined in the previous section can be extended in this notation for the linear methods (left) and the non-linear methods (right) as:

Upwind	$\phi(\theta) = 0$	MinMod	$\phi(\theta) = minmod(1, \theta)$
Lax-Wendroff	$\phi(\theta) = 1$	SuperBee	$\phi(\theta) = max(0, min(1, 2\theta), min(2, \theta))$
Beam-Warming	$\phi(\theta) = \theta$	MC	$\phi(\theta) = max(0, min((1+\theta)/2, 2, \theta)$
Fromm	$\phi(\theta) = 0.5(1+\theta)$	van Leer	$\phi(\theta) = 0.5(1+\theta)$

If we apply these limiters to the advection equation which was discussed in Figure 2, it can be seen that the limiters help to capture the discontinuity sharper.

For the most of the limiters, it is not clear to see whether they are a TVD scheme or not, therefore a formal definition to check the TVD property. The theorem of Harten states that

Theorem 1: A general method which is written in the form

$$Q^{n+1}_i = Q^n_i - C^n_{i-1}(Q^n_i - Q^n_{i-1}) + D^n_i(Q^n_{i+1} - Q^n_i)$$

is TVD, if all the following conditionsa are satisfied:

- $C^n_{i-1} \geq 0$
- $D^n_i \geq 0$
- $C^n_i + D^n_i \leq 1$

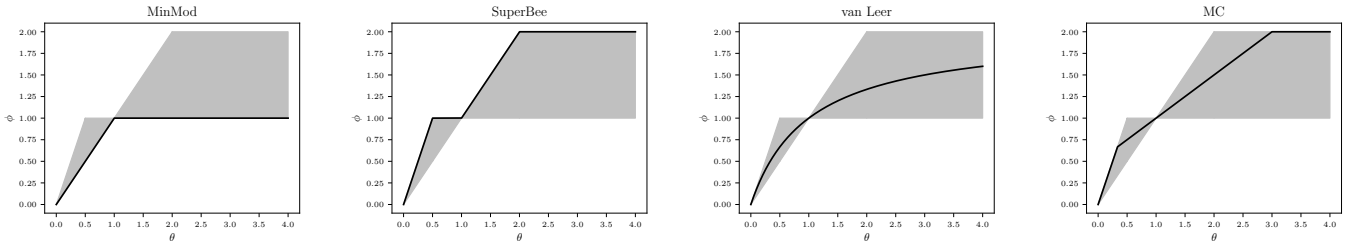

Fig. 3. The TVD region in $\phi - \theta$ space and the locations of the TVD limiters

and writing the corresponding terms as (using $\nu = \bar{u}\Delta t/\Delta x$):

$$C_{i-1}^n = \nu + \frac{1}{2}\nu(1-\nu)\left(\frac{\phi(\theta_{i+1/2}^n)}{\theta_{i+1/2}^n} - \phi(\theta_{i-1/2}^n)\right)$$

$$D_i^n = 0$$

Then the TVD condition is satisfied when:

$$0 \leq C_{i-1}^n \leq 1$$

Finally, using the derived relation for the C_{i-1}^n and restricting the CFL number $\nu = \bar{u}\Delta t/\Delta x$ to $0 \leq \nu \leq 1$, relation for the θ and $\phi(\theta)$ can be written as:

$$\theta \leq \phi(\theta) \leq 2\theta, (0 \leq \theta \leq 1)$$

$$1 \leq \phi(\theta) \leq \theta, (1 \leq \theta \leq 2)$$

$$1 \leq \phi(\theta) \leq 2, (\theta > 2)$$

This relation defines a region in the $\theta - \phi(\theta)$ space, which is shown in the Figure 3 as the gray area, which is also called as the Sweby diagram. The TVD limiters are also shown in the figure. The MinMod limiter lies on the lower bound of the region, since it chooses the lower slope. On the other hand, the SuperBee limiter lies on the upper bound, as opposed to MinMod. MC limiter follows more conservative approach and has a smooth relation in the bottleneck region. van Leer is the most different one among the 4, since it is the only smooth limiter on the entire region. In the literature there are lot more limiters which are developed such that they fit in the corresponding region.

V. High Resolution in the Temporal Domain

Until now, the developed methods were discrete in the both temporal and spatial domains. If we discretize a PDE in the temporal space first, it reduces the PDE to a system of ODEs, which can be solved using the known ODE solvers. This approach allows us to introduce higher order methods both in the spatial and the temporal domain, with decoupled issues of the discretizations of each domain. We can write the resulting semi-discrete version of the update formula as:

$$Q_i'(t) = -\frac{1}{\Delta x}[F_{i+1/2}(Q(t)) - F_{i-1/2}(Q(t))] = \mathcal{L}_i(Q(t))$$

In this form, spatial and temporal discretizations are decoupled. For the spatial discretization, we can apply the previous knowledge about limiters and high-order methods. Then, we can apply more advanced methods for the time discretization. However, as in the spatial discretization, a special care must be taken while constructing the higher order methods in the temporal domain due to the possibility of the additional oscillations. In order to ensure that we do not add oscillations by our numerical methods, we again have to make sure that the method shows the TVD characteristic. TVD methods based on the semi discretization approach are easy to verify. If a spatial discretization method is a TVD method when the forward Euler time stepping is used, some of the high-order time stepping methods is guaranteed to be a TVD method too. These kind of time stepping methods are called strong stability-preserving (SSP) time discretizations. An example of one of these special methods is the two step Runge-Kutta time stepping:

$$Q^* = Q^n + \Delta t\mathcal{L}(Q^n)$$

$$Q^{**} = Q^* + \Delta t\mathcal{L}(Q^*)$$

$$Q^{n+1} = \frac{1}{2}(Q^n + Q^{**})$$

VI. ENO Schemes

For the high order reconstruction, we have chosen to switch from the piecewise constant polynomial to a piecewise linear polynomial. Naturally, one might think how to increase the degree of the polynomial even further. However, the problem is that we only have cell averages to construct the polynomial and if we try to increase the degree of the polynomial, number of free parameters would be much higher. Instead of approaching it blindly, consider a function such that:

$$w'(x) = q(x,t)$$

and

$$w(x) = \int_{x_{1/2}}^{x} q(\xi,t)d\xi$$

where the lower bound of the integral is arbitrary. For example, a cell average can be used. Then we can write

$$W_i = w(x_{i+1/2}) = \int_{x_{1/2}}^{x_{i+1/2}} q(\xi,t)d\xi$$

Then, given the cell averages, this function is going to yield the summation of the cell averages

$$W_i = \Delta x \sum_{j=1}^{i} \bar{q}_j(t)$$

Basically, to approximate the w in the cell i, we can use an interpolating polynomial of degree s, passing through $s+1$ points.

The flaw of this approach is that the cell averages should be smooth in those $s+1$ points. However, it is not the case for the most of the problems, especially the non-linear ones. Consider that we interpolate the values $W_{i-j}, ..., W_{i-j+s}$. High degree interpolant polynomials are highly oscillatory even for a smooth data. We can adapt the idea to minimize the oscillations, which was introduced in the limiter based methods, to this concept too. We need to choose the value j for each i such that the interpolating polynomials result in the smallest possible oscillations.

To develop such a set of interpolant polynomials, one can use the Newton's divided difference method. Starting with the linear function of passing W_{i-1} and W_i, one can write higher order functions recursively. Note that the divided differences of W_i are directly related to the values $\bar{q}_i$ too. Therefore, the values for W_i are not calculated in practice. With this approach, one can obtain arbitrary degree of polynomials depending on the smoothness of the region with high efficiency. This type of methods are called essentially non-oscillatory (ENO) methods. Another popular version of the vanillia ENO method is the weighted ENO (WENO) in which all the divided differences are used to calculate a weighted average. Since they are already calculated, this approach does not result in additional computational effort. Moreover, since this approach uses all the differences in the weighted manner, it has the capability to adapt itself in the smooth and discontinuous regions better than the vanilla version, which results in more robust method.

VII. Conclusion

The higher-order methods for the conservation laws needs special attention due to their close relationship with the discontinuous solutions. Although the linear equations cannot result in discontinuous solutions with a smooth initial data, non-linear equations or linear equations with discontinuous initial data can. Therefore, the methods developed here mostly considered the discontinuous initial data on a linear conservation law, due to its simplicity and their representational power. These methods are easily extensible to non-linear equations too.

The first approach, the Lax-Wendroff method, was to obtain a higher-order method using Taylor series expansion, which introduced non-physical oscillations to the solution. Then, it is improved using limiters to detect the smooth and discontinuous regions in the solution, and the corresponding TVD methods were discussed. Since we also want higher resolution in the temporal domain too, the semi-discrete approach was discussed. Finally, more advanced methods, which are based on the higher order polynomial reconstruction, ENO and WENO methods are developed. These are the main recipe for the higher-order methods and more advanced methods can be found on the literature with this background information.

Appendix A
Implementations of the Methods

A simple Python code to solve linear advection equation and Burger's equation (experimental) with several limiters can be found in *github.com/oguzziya/high-order-fv*.

References

[1] R.J. Leveque, *Finite-Volume Methods for Hyperbolic Problems*, 2nd ed. Cambridge, England: Cambridge University Press, 2004.
[2] E.F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamic, A Practical Introduction*, 3rd ed. Heidelberg, Germany: Springer, 2009.
[3] T.J. Barth and H. Deconinck, *High-Order Methods for Computational Physics*, 1st ed. Heidelberg, Germany: Springer, 1999.