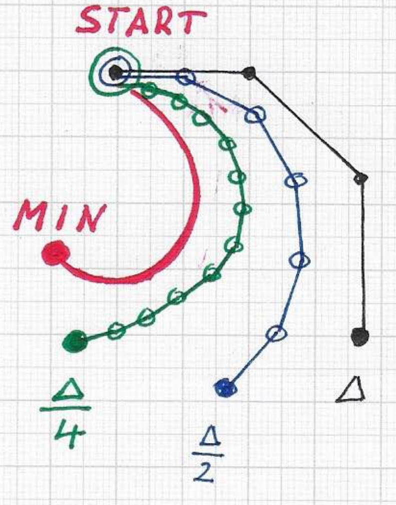
Bernd Hamann
Stratovan

3/10/2021

OPTIMIZATION AND RICHARDSON EXTRAPOLATION

• **WHY?** Minimizing a Loss (or cost) function — depending on "weights" to be optimized in a high-dimensional space — is fundamental for artificial neural networks. Minimization should be 'robust' and as efficient as possible (Gradient descent, improved 'Nestero gradient descent', Line search and stochastic gradient descent methods are examples of typically used minimization appraaches.)

• **IDEA:** (Considering computation/approximation of integral/tangent curves of a GRADIENT VECTOR FIELD for Loss function minimization)



0) A gradient descent method computes a set of points converging (under certain conditions) to the minimum, using as start point a point close to the minimum.

1) A constant or an adaptive stepsize is used to move from point to point.

2) A gradient vector field's integral/tangent curves can be computed/approximated using EULER's simple method, for example.

"Sequence of polylines" converging to exact curve (red) for stepsizes Δ, $\Delta/2$, $\Delta/4$. $\Delta \to 0$
$\Rightarrow$ convergence to exact curve.

OPTIMIZATION AND RICHARDSON EXTRAPOLATION – Cont'd.

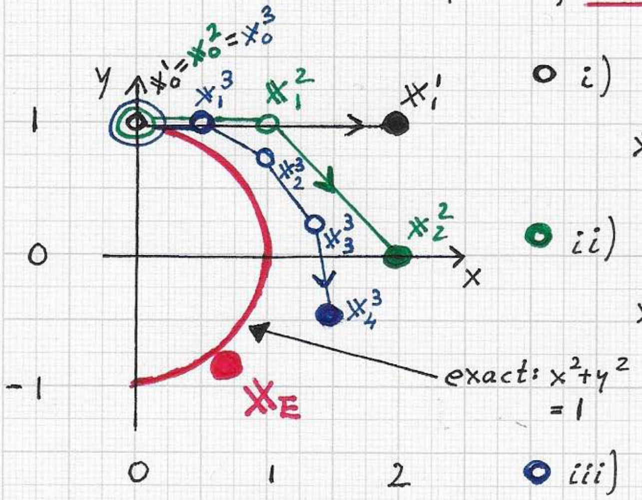
... 3) Compute several polyline approximations, all beginning at the same start point ("time=0") and ending at different terminal points ("time = T").
To simplify the explanation of the principle, use stepsizes ("time steps") Δ, $\Delta/2$, $\Delta/4$, ... for the computations of the points of the polylines.
FOR THE PURPOSE OF LOSS FUNCTION MINIMIZATION, ONLY THE TERMINAL POINTS OF THE POLYLINES MATTER; THEY ARE ALL VIEWED AS "NEAR-MINIMUM POINTS." RICHARDSON EXTRAPOLATION IS APPLIED TO THE SET OF TERMINAL POINTS (I.E., $\Delta \to 0$) TO OBTAIN AN IMPROVED MINIMUM APPROXIMATION.

- Ex.: To explain the principle of the use of Richardson extrapolation applied to terminal point improvement for 3 polylines obtained with stepsizes Δ, $\Delta/2$, $\Delta/4$, consider the vector field with circles as tangent curves: $\mathbb{V} = \mathbb{V}(x,y) = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix} = 2\begin{pmatrix} y \\ -x \end{pmatrix}$. (This vector field is a field orthogonal to the gradient vector field of $f(x,y) = x^2+y^2$, having circles as its isocurves.) Use the simple EULER scheme $\mathbb{X}_{i+1} = \mathbb{X}_i + \mathbb{V}_i \cdot \Delta = \mathbb{X}_i + \Delta \cdot \mathbb{V}(\mathbb{X}_i)$ to compute points.

Stratovan

■ OPTIMIZATION AND RICHARDSON EXTRAPOLATION – Cont'd

...• Ex.: Compute 3 polylines for $\begin{pmatrix} u \\ v \end{pmatrix} = 2\begin{pmatrix} y \\ -x \end{pmatrix}$, using $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ as start point; compute 2, 3 and 5 points, respectively.



o i) Polyline 1, Δ:

$$x_0^1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad x_1^1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \Delta \cdot 2 \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2\Delta \\ 1 \end{pmatrix}$$

o ii) Polyline 2, $\Delta/2$:

$$x_0^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad x_1^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \frac{\Delta}{2} \cdot 2 \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \Delta \\ 1 \end{pmatrix}$$

$$x_2^2 = \begin{pmatrix} \Delta \\ 1 \end{pmatrix} + \frac{\Delta}{2} \cdot 2 \cdot \begin{pmatrix} 1 \\ -\Delta \end{pmatrix} = \begin{pmatrix} 2\Delta \\ 1-\Delta^2 \end{pmatrix}$$

o iii) Polyline 3, $\Delta/4$:

$$x_0^3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad x_1^3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \frac{\Delta}{4} \cdot 2 \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \Delta/2 \\ 1 \end{pmatrix}$$

$$x_2^3 = \begin{pmatrix} \Delta/2 \\ 1 \end{pmatrix} + \frac{\Delta}{4} \cdot 2 \begin{pmatrix} 1 \\ -\Delta/2 \end{pmatrix} = \begin{pmatrix} \Delta \\ 1-\Delta^2/4 \end{pmatrix}$$

$$x_3^3 = \begin{pmatrix} \Delta \\ 1-\Delta^2/4 \end{pmatrix} + \frac{\Delta}{4} \cdot 2 \begin{pmatrix} 1-\Delta^2/4 \\ -\Delta \end{pmatrix} = \begin{pmatrix} \Delta + \Delta/2\,(1-\Delta^2/4) \\ 1-\Delta^2/4 - \Delta^2/2 \end{pmatrix} = \begin{pmatrix} \Delta + \Delta/2 - \Delta^3/8 \\ 1 - 3/4\,\Delta^2 \end{pmatrix} = \frac{1}{8}\begin{pmatrix} \Delta(12-\Delta^2) \\ 8-6\Delta^2 \end{pmatrix}$$

$$x_4^3 = \ldots = \begin{pmatrix} \Delta(2-\Delta^2/2) \\ 1 - 3/2\,\Delta^2 + \Delta^4/16 \end{pmatrix}$$

⇒ The 3 terminal points of the 3 polylines are x_1^1, x_2^2, x_4^3.

⇒ Richardson extrapolation can be applied to the 3 terminal points.

• Ex.: $\Delta = 1$ ⇒

i) $x_0^1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, x_1^1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ = ●

ii) $x_0^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, x_1^2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, x_2^2 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$ = ●

iii) $x_0^3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, x_1^3 = \begin{pmatrix} 1/2 \\ 1 \end{pmatrix}, x_2^3 = \begin{pmatrix} 1 \\ 3/4 \end{pmatrix}$,

$x_3^3 = \begin{pmatrix} 11/8 \\ 1/4 \end{pmatrix}, x_4^3 = \begin{pmatrix} 3/2 \\ -7/16 \end{pmatrix}$ = ●

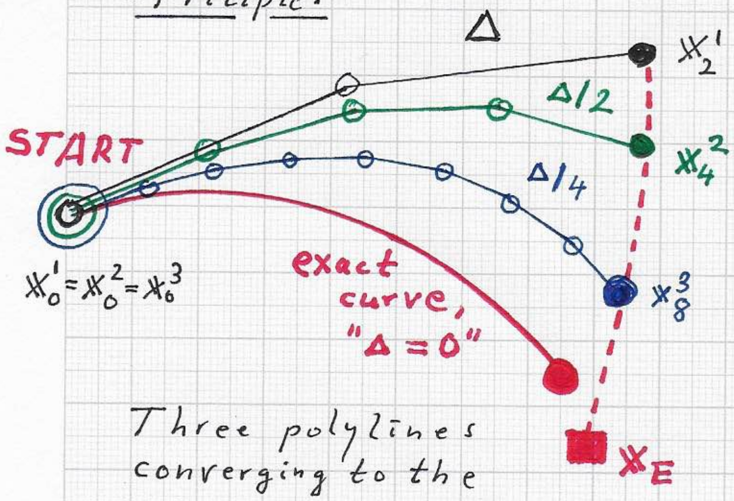
⇒ Richardson extrapolation for these 3 terminal points:

$$x_E = \left(8 \cdot \begin{pmatrix} 3/2 \\ -7/16 \end{pmatrix} - 6 \cdot \begin{pmatrix} 2 \\ 0 \end{pmatrix} + 1 \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix}\right) / 3 = \begin{pmatrix} 2/3 \\ -5/6 \end{pmatrix} = ●$$

Stratovan

▪ OPTIMIZATION AND RICHARDSON EXTRAPOLATION – Cont'd.

• Priciple:



Three polylines converging to the exact tangent curve

→ Construct multiple polyline approximations of exact tangent curve – "terminating very close to minimum"

→ All polyline approximations have a terminal points associated with the same "overall time T".

→ Apply Richardson extrapolation to set of terminal points $x_2^1, x_4^2, x_8^3, x_{16}^4, \ldots$ obtained with polyline stepsizes $\Delta, \Delta/2, \Delta/4, \Delta/8, \ldots$

(→ Terminal points = "good" approximations of minimum)

⇒ x_E is the extrapolated point that represents the final approximation of the desired minimum.

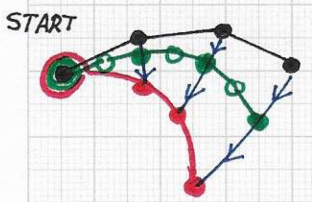
• Notes: i) To compute points of a polyline approximation of an integral curve of a gradient vector field, higher-order Runge-Kutta schemes (with adaptive stepsize) should be used. ("Runge-Kutta-order-4 with adaptive stepsize"!)

ii) It is important that the individual polylines can be "associated with" a meaningful, average stepsize $\Delta, \Delta/2, \Delta/4, \Delta/8, \ldots$ to make possible Richardson extrapolation via the terminal points. ~BH

OPTIMIZATION AND RICHARDSON EXTRAPOLATION – Cont'd.

• **Notes:** iii) In the case of minimizing a loss function that depends on many parameters (e.g., weights needed to 'optimally' combine node output data in a neural network), one would generate a small number of gradient vector field tangent curve approximations — all to end 'close' to a minimum in parameter space. The terminal points of these tangent curve approximations must be associated with 'time steps' Δ, $\Delta/2$, $\Delta/4$, $\Delta/8$, ..., and the final value of the minimum is the result of Richardson extrapolation applied to terminal points.

iv) To determine the minimum of a (loss) function, Richardson extrapolation is applied only to the terminal points of multiple tangent curve approximations ending closely to the minimum; it is possible to extrapolate all points of multiple tangent curve approximations – by extrapolating with points of these approximations having the same 'time' (= sum of total stepsizes used) associated with them. This possibility can be considered to 'explode', to expand, to extrapolate the region (or region boundary) in feature space defining a (material) class.



{•} Δ points

{•} $\frac{\Delta}{2}$ points

{•} $\Delta = 0$ points

⇒ Richardson extrapolation generates ($\Delta = 0$) pnts., defining a high-quality tangent curve approximation.