A Target Oriented Averaged Search Trajectory and its Applications in Artificial Neural Networks

Ángel Rojas
Ph.D. Andreas Griewank

Yachay Tech

28th May 2019

Outline

- Artificial Neural Networks (ANN)
  - Optimization task and learning algorithms
  - The optimizer's house of horror.
- Global Optimization (Nonsmooth)
  - Target Oriented Average Search Trajectory (TOAST)
  - Successive Abs-Linear Global Optimization (SALGO)
- Results and comparison
- Conclusions

Artificial Neural Networks (ANN)

"Machine Learning is the science (and art) of programming computers so they can learn from data." [1]
ANN is a data-based model in order to predict data on basis of previous training on similar data.
Such a model is called prediction function to determine an empirical risk measure based on training data.

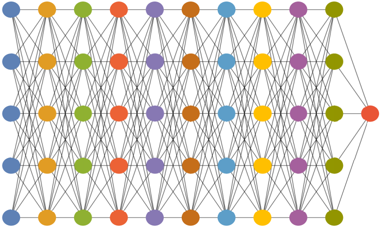

Figure 1: A fully-connected-Artificial Neural Network

Optimization task and learning algorithms

$$\min_{W} \phi(W) \equiv \frac{1}{m} \sum_{k=1}^{m} |f(W, x_k) - y_k|$$

over a training set of m pairs $(x_k, y_k) \in \mathbb{R}^{n+1}$

Learning Algorithms

- Steepest Descent, i.e., Backpropagation
- Gradient Momemtum Variants.
- Stochastic Gradient Method

Specially, for SG choice of stepsize is crucial but very difficult.
Our concern is the efficiency of the optimization procedure on the training data rather than analyzing how the model is efficient for testing.

House of Horrors

A single-layer case with constant output weighting $p \in \{-1, 1\}^d$ and hinge activation (ReLU) can be mathematically described by the predictor:

$$f(W, x) \equiv p^\top \max(0, W_{1\dots n}x + W_{n+1}) \text{ with } W \in \mathbb{R}^{d(n+1)}$$

- **Nonsmoothness**
  At all isolated local and at least one global optimizer $\phi(W)$ is not differentiable.
- **Multi-modality**
  There may be local minima with values high above the globally minimal value.
- **Singularity.**
  In the multi-layer case $f(W)$ is constant along invariant paths $W(t)$ even at stationary points.
- **Zero-PLateau**
  For large negative W_{n+1} the function $f(W, x)$ and the gradient $\nabla\phi(W)$ w.r.t. W and x vanish identically.

Example with two variable weights

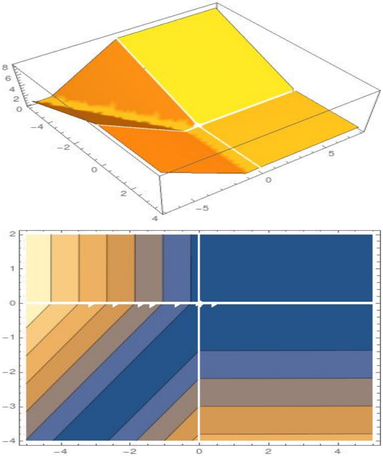

Figure 2: One-layer ANN model and its contours

Global Optimization (Nonsmooth)

Most optimization methods move down hill to reach a local minimizer or possibly a saddle point.
To find the smallest of these local minimizers x_* is generally a very difficult problem.

$$\varphi(x_*) \leq \varphi(x), \forall x \in \mathcal{D}$$

Space covering techniques

If $x \in \mathbf{R}^n, n \geq 2$, these methods tend to exceed computational limitation as they have to sample the function on a set of points that is sufficiently dense to cover the search area.

Non-rigorous techniques

- Stochastic/Statistics-based searches
- Deterministic, but heuristic searches (many parameters).
- Hybrid methods

Target Oriented Average Search Trajectory (TOAST)

$$\ddot{x}(t) = -\left(I - \frac{\dot{x}(t)\dot{x}(t)^{\top}}{\|\dot{x}(t)\|^2}\right)\frac{\nabla\phi(x(t))}{[\phi(x(t)) - c]}, \text{ with } \|\dot{x}(t_0)\| = 1$$

- Idea: Adjustment of current search direction $\dot{x}(t)$ towards the steepest descent direction.
- The closer the current function value $\phi(x(t))$ is to the target level c, the more rapidly the direction is adjusted.
- In the limit when $\phi(x(t))$ tends to c the trajectory reduces to steepest descent.
- On homogeneous objectives, local minimizers below c are accepted and local minimizers above the target level are passed by.

Closed form solution on prox-linear function

Theorem. If $\varphi(x) = g^\top x + b + \frac{q}{2}\|x\|_2^2$

$$\ddot{x}(t) = -\left[I - \dot{x}(t)\,\dot{x}(t)^\top\right] \frac{\nabla\varphi(x(t))}{[\varphi(x(t)) - c]}$$

implies

$$x(t) = x_0 + \frac{\sin(\omega t)}{\omega}\dot{x}_0 + \frac{1-\cos(\omega t)}{\omega^2}\ddot{x}_0 \tag{1}$$

and

$$\varphi(x(t)) \;=\; \varphi_0 + \left[(g + q x_0)^\top \dot{x}_0\right] \frac{\sin(\omega t)}{\omega} + \left[q - \omega^2(\varphi_0 - c)\right] \frac{(1-\cos(\omega t))}{\omega^2} \tag{2}$$

where

$$\ddot{x}_0 = -\left[I - \dot{x}_0\dot{x}_0^\top\right] \frac{(g + q x_0)}{(\varphi_0 - c)} \quad \text{and} \quad \omega = \|\ddot{x}_0\|\,. \tag{3}$$

Theorem[3],[4]

1. Every function $\varphi(x)$ that is evaluated by a sequence of smooth elemental functions and piecewise linear elements like abs, min, max can be approximated near a reference point $\mathring{x}$ by a piecewise-linear function $\Delta\varphi(\mathring{x};\Delta x)$ s.t.

$$|\varphi(\mathring{x}+\Delta x)-\varphi(\mathring{x})-\Delta\varphi(\mathring{x};\Delta x)|\leq \tfrac{q}{2}\|\Delta x\|^2$$

2. The function $y=\Delta\varphi(\mathring{x};x-\mathring{x})$ can be represented in Abs-Linear form

$$\begin{aligned} z &= d+Zx+Mz+L|z|, \\ y &= \mu+a^\top x+b^\top z+c^\top|z| \end{aligned}$$

where Z and L are strictly lower triangular matrices s.t. $z=z(x)$.

This form can be generated automatically by Algorithmic Differentiation and it allows the computational handling of $\Delta\varphi$ in and between the polyhedra

$$P_\sigma = cl\{x\in\mathbb{R}^n;\mathrm{sgn}(z(x))=\sigma\}$$

SALGO-TOAST algorithm

1. Form piecewise linearization $\Delta\varphi$ of objective φ at the current iterate $\mathring{x}$ and estimate the proximal coefficient q, set $x_0 = \mathring{x}$,
2. Select the initial tangent $\dot{x}_0$ and $\sigma = \operatorname{sgn}(z(x_0))$.
3. Compute and follow circular segment $x(t)$ in P_σ.
4. Determine minimal t_* where $\varphi(x(t_*)) = c$ or $x_* = x(t_\star)$ lies on the boundary of P_σ with some $P_{\tilde{\sigma}}$.
5. If $\varphi(x_*) < c$, lower c or go to step (1) with $\mathring{x} = x_*$ or terminate.
6. Else, set $x_0 = x_*$, $\dot{x}_0 = \dot{x}(t_*)$, $\sigma = \tilde{\sigma}$ and continue with step (3).

TOAST path

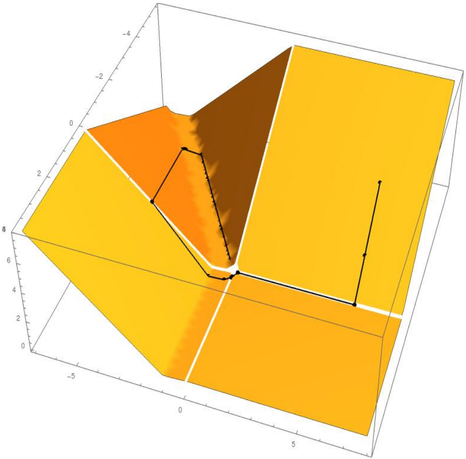

Figure 3: Reached minimum value 0.591576 and target level 0.519984

Griewank function in 2D with 10 intermediate nodes and 20 training data points

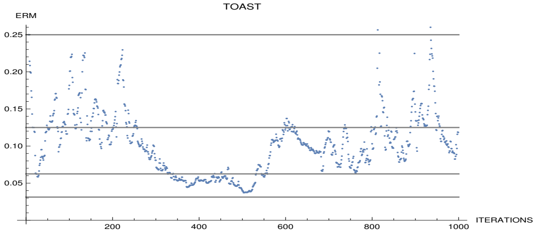

Figure 4: TOAST-SALGO with minimum 0.037252 and target level 0.031233

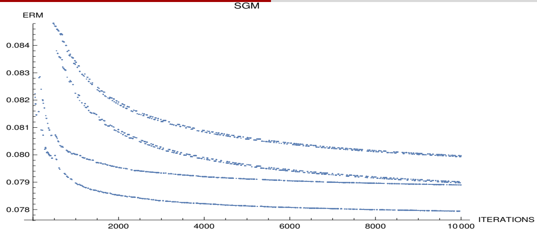

Figure 5: Stochastic Gradient Method implementation with minimum 0.077943

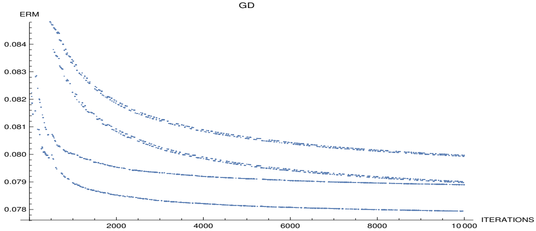

Figure 6: Gradient descent implementation

Remain Tasks and further development

1. Refining targeting and restarting strategy.
2. Extension to ”deep learning”
3. Application to standard problem MNIST
4. Matrix based implemmentation for HPC
5. Explotation of low-rank updates in polyhedral transition.
6. Sample-wise version in Stochastic Gradient fashion

References

- A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.", 2017.

- A. Griewank, "On stable piecewise linearization and generalized algorithmic differentiation," *Optimization Methods and Software*, vol. 28, no. 6, pp. 1139–1178, 2013.

- A. Griewank and A. Walther, "Finite convergence of an active signature method to local minima of piecewise linear functions," *Optimization Methods and Software*, pp. 1–21, 2018.

- A. Griewank, A. Walther, and P. N. SPP1962, "Priority programme 1962."

- Thank You.

Introduction and Motivation

- Artificial Neural Network yields nonsmooth and, in general, nonconvex functions w.r.t. weights, shifts, and input data.
- These functions can be written in Abs-Normal Form (ANF) and, consequently, Abs-Linear Form (ALF). The latter has a uniform proximal quadratic term $\|\frac{q}{2}\Delta x\|^2, q > 0$ w.r.t. original model.
- Nonsmooth optimality conditions are NP-hard to satisfy and there is no stopping criteria in the nonconvex case.
- A common used ANN activation function is hinge function (a.k.a. ReLU), a suitable piecewise-linear function for ANF.
- Formulation of a global nonsmooth optimization method based on a Target Oriented Average Search Trajectory and Successive Abs-Linearization routine, namely, TOAST and SALGO, respectively.

Tentative comparison

- TOAST-SALGO achieves lower minima than SGM and GD implementations
- SGM and GD seems to get stuck in local minima, i.e., zigzagging and V-shaped valley.
- TOAST-SALGO solves the zig-zagging problem, climbing up and rolling down to achive a new target level.
- The singularities of gradient and Hessian is a problematic in SGM and GD.