Generalized/Global Abs-Linear Learning (GALL)

Andreas Griewank and Ángel Rojas

Humboldt University (Berlin) and Yachay Tech (Imbabura)

14.12.19, NeurIPS Vancouver

Outline

Folklore and Common Expectations in ML

1. Nonsmoothness can be ignored except for step size choice.
2. Stochastic (mini-batch) sampling hides all the problems.
3. Higher dimensions make local minimizer less likely.
4. Difficulty is getting away from saddle points not minimizers.
5. Precise location of (almost) global minimizer unimportant.
6. Network architecture and stepsize selection can be tweaked.
7. Convergence proofs only under "unrealistic assumptions".

Generalized Gradient Concepts

Notational Zoo (Subspecies in Euclidean and Lipschitzian Habitat):

Fréchet Derivative: $\nabla\varphi(x) \equiv \partial\varphi(x)/\partial x : \mathcal{D} \mapsto \mathbb{R}^n \cup \emptyset$

Limiting Gradient: $\partial^L f(\mathring{x}) \equiv \overline{\lim}_{x \to \mathring{x}} \nabla\varphi(x) : \mathcal{D} \rightrightarrows \mathbb{R}^n$

Clarke Gradient: $\partial\varphi(x) \equiv \mathbf{conv}(\partial^L\varphi(x)) : \mathcal{D} \rightrightarrows \mathbb{R}^n$

Bouligand: $f'(x;\Delta x) \; \equiv \; \lim_{t \searrow 0}[\varphi(x + t\Delta x) - \varphi(x)]/t$

$$: \quad \mathcal{D} \times \mathbb{R}^n \mapsto \mathbb{R}$$

$$: \quad \mathcal{D} \mapsto \mathbf{PL}_h(\mathbb{R}^n)$$

Piecewise Linearization(PL):

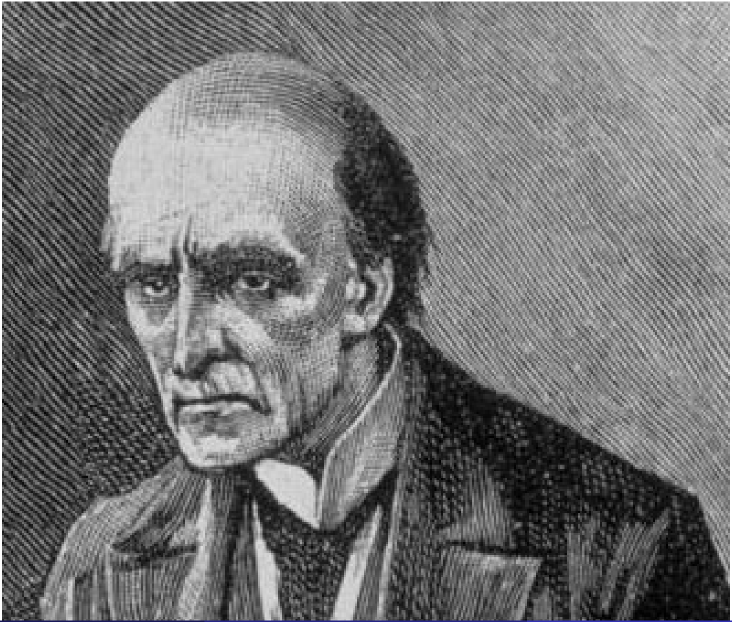
$$\Delta\varphi(x;\Delta x) \quad : \quad \mathcal{D} \times \mathbb{R}^n \mapsto \mathbb{R}$$

$$: \quad \mathcal{D} \mapsto \mathbf{PL}(\mathbb{R}^n)$$

Moriarty Effect due to Rademacher ($C^{0,1} = W^{1,\infty}$) :

Almost everywhere all concepts reduce to Fréchet, except PL!!

Lurking in the background: Prof. Moriarty



Filippov solutions of generalized steepest descent inclusion

The convexity and outer semi-continuity of subsets $\partial\varphi(x(t))$ imply that

$$-\dot{x}(t) \;\in\; \partial\varphi(x(t)) \quad \text{from} \quad x(0) = x_0 \in \mathbb{R}^n$$

has (at least) one absolutely continuous Filippov solution trajectory $x(t)$.

Heavy ball (Polyak, 1964)

$$-\ddot{x}(t) \;\in\; \partial\varphi(x(t)) \quad \text{from} \quad x(0) = x_0, \; -\dot{x}(0) \in \partial\varphi(x_0)\;.$$

Picks up speed/momentum going downhill and slows down going uphill.

Savvy ball (Griewank, 1981)

$$\frac{d}{dt}\left[\frac{-\dot{x}(t)}{(\varphi(x(t))-c)^e}\right] \;\in\; \frac{e\,\partial\varphi(x(t))}{(\varphi(x(t))-c)^{e+1}} \;=\; \partial\left[\frac{-1}{(\varphi(x(t))-c)^e}\right]\;.$$

Can be rewritten as a first order system of a differential equation and an inclusion satisfying Fillipov $\implies$ absolutely continuous $(x(t), \dot{x}(t))$ exists.

Integrated Form

$$v(t) = \frac{\dot{x}(t)}{[\varphi(x(t)) - c]^e} \in \frac{\dot{x}_0}{[\varphi(x_0) - c]^e} - e \int_0^t \frac{\partial\varphi(x(\tau))\, d\tau}{[\varphi(x(\tau)) - c]^{e+1}} \, .$$

Second order Form

$$\ddot{x}(t) \in - \left[I - \frac{\dot{x}(t)\, \dot{x}(t)^\top}{\|\dot{x}(t)\|^2} \right] \frac{[e\, \partial\varphi(x(t))]}{[\varphi(x(t)) - c]} \quad \text{with} \quad \|\dot{x}(0)\| = 1 \, .$$

- Idea: Adjustment of current search direction $\dot{x}(t)$ towards a negative gradient direction $-\partial\varphi(x(t))$.
- The closer the current function value $\varphi(x(t))$ is to the target level c, the more rapidly the direction is adjusted.
- If φ convex, $\varphi(\mathring{x}) \leq c$ and $e \leq 1$ the trajectory reaches the level set.
- On degree $(1/e)$ homogeneous objectives, local minimizers below c are accepted and local minimizers above the target level are passed by.

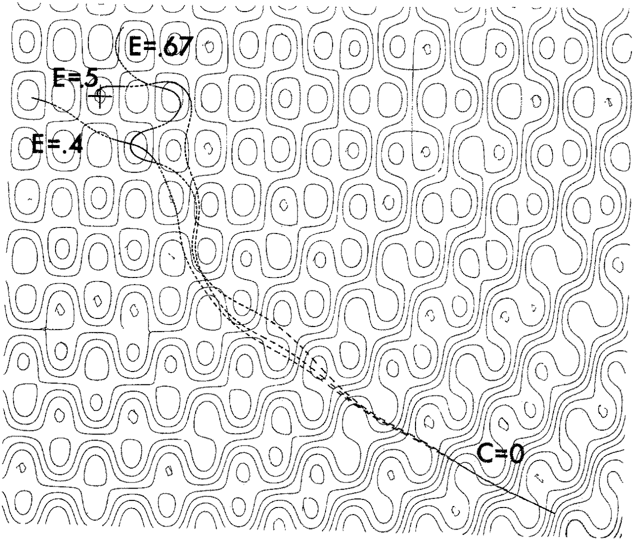


Fig. 1. Search trajectories with target $c = 0$ and sensitivity $e \in \{0.4, 0.5, 0.67\}$ on the objective function $f = (x_1^2 + x_2^2)/200 + 1 - \cos x_1 \cos(x_2/\sqrt{2})$. Initial point $(40, -35)$. Global minimum at origin marked by $+$.

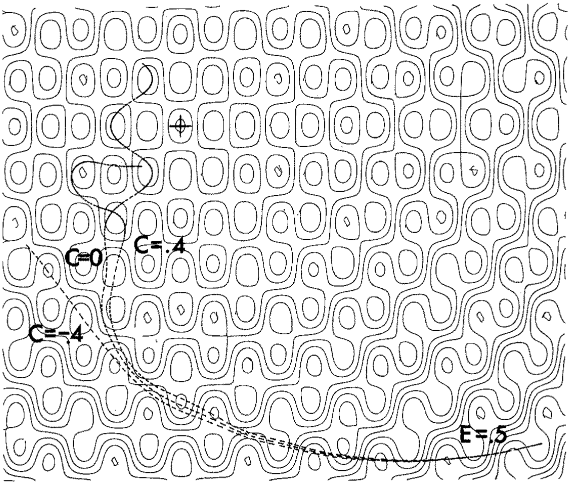


Fig. 2. Search trajectories with sensitivity $e = 0.5$ and target $c \in \{-0.4, 0, 0.4\}$ on the objective function $f = (x_1^2 + x_2^2)/200 + 1 - \cos x_1 \cos(x_2/\sqrt{2})$. Initial point $(35, -30)$. Global minimum at origin marked by +.

Closed form solution on prox-linear function

Lemma(A.G. 1977 & A.R. 2019). For $\varphi(x) = b + g^\top x + \frac{q}{2}\|x\|_2^2$

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

yields momentum like

$$x(t) = x_0 + \tfrac{\sin(\omega t)}{\omega}\dot{x}_0 + \tfrac{1-\cos(\omega t)}{\omega^2}\ddot{x}_0 \;\approx\; x_0 + t\dot{x}_0 - \tfrac{t^2 g}{2(\varphi_0 - c)}$$

and

$$\varphi(x(t)) \;=\; \varphi_0 + \left[(g + qx_0)^\top \dot{x}_0\right] \tfrac{\sin(\omega t)}{\omega} + \left[q - \omega^2(\varphi_0 - c)\right] \tfrac{(1-\cos(\omega t))}{\omega^2}$$

where

$$\ddot{x}_0 = -\left[I - \dot{x}_0\dot{x}_0^\top\right] \frac{(g + qx_0)}{(\varphi_0 - c)} \quad \text{and} \quad \omega = \|\ddot{x}_0\|\,.$$

Piecewise-Linearization Approach

1. Every function $\varphi(x)$ that is *abs-normal*, i.e. 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.

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

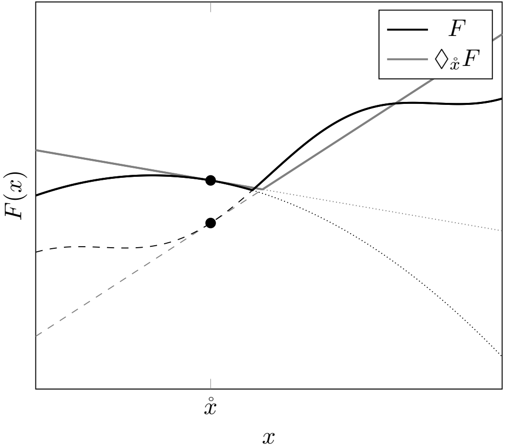
2. The function $y = \Delta\varphi(\mathring{x}; x - \mathring{x})$ can be represented in Abs-Linear form

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

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

3. $[d, Z, M, L, \mu, a, b, c]$ can be generated automatically by Algorithmic Piecewise Differentiation, which allows the computational handling of $\Delta\varphi$ in and between the polyhedra

$$P_\sigma = closure\{x \in \mathbb{R}^n; \operatorname{sgn}(z(x)) = \sigma\} \quad \text{for} \quad \sigma \in \{-1, +1\}^s$$



(a) Tangent mode linearization

SALMIN defined by iteration

$$x_{k+1} \ = \ \underset{\Delta x}{\text{arglocmin}}\{\Delta\varphi(x_k;\Delta x) + \ \tfrac{q_k}{2}\|\Delta x\|^2\} \tag{1}$$

where $q_k > 0$ is adjusted such that eventually $q_k \geq q$ in region of interest. Has cluster points x_* that are first order minimal minimal (FOM) i.e.

$$\Delta\varphi(x_*,\Delta x) \ \geq \ 0 \quad \text{for} \quad \Delta x \approx 0\,.$$

Drawback: Requires computation and factorization of active Jacobians.

Coordinate Global Descent **CGD**

$f(w;x)$ is PL w.r.t. x but $\varphi(w)$ is only multi-piecewise linear w.r.t. w, i.e.

$$\varphi(x+te_j) \ \equiv \ \ \varphi(x) \ + \ \Delta\varphi(x+te_j) \quad \text{for} \quad t\in\mathbb{R} \quad .$$

Along any such coordinate bi-direction we can perform a global univariate minimization efficiently. Cluster points x_* of such alternating coordinate searches seem not even even Clarke stationary, i.e. $0 \in \partial\varphi(x_*)$.

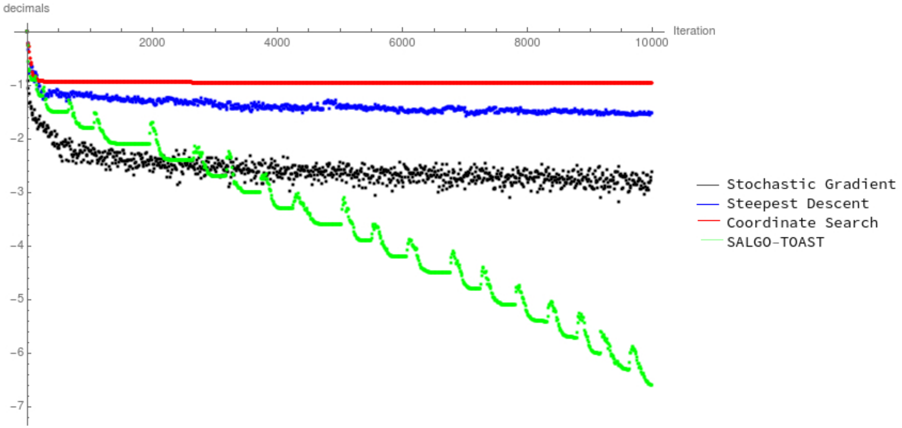


Figure 1: Decimal digits gained by 4 methods on single layer regression problem.

SALGO-SAVVY 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 $\sigma = \text{sgn}(z(x_0))$.
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 $\varphi(x_*) \leq c$ then lower c and goto step (2) // restart inner loop xor goto step (1) with $\mathring{x} = x_*$ and adjusted q // continue outer loop xor terminate optimization if user "happy" or resources exceeded.
6. Else, set $x_0 = x_*$, $\dot{x}_0 = \dot{x}(t_*)$, $\sigma = \tilde{\sigma}$ and continue with step (3).

Many other heuristic strategies for retargeting and restarting possible.!!!!

Savvy Ball Path

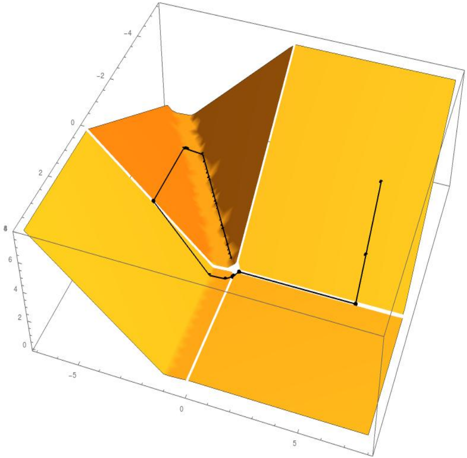
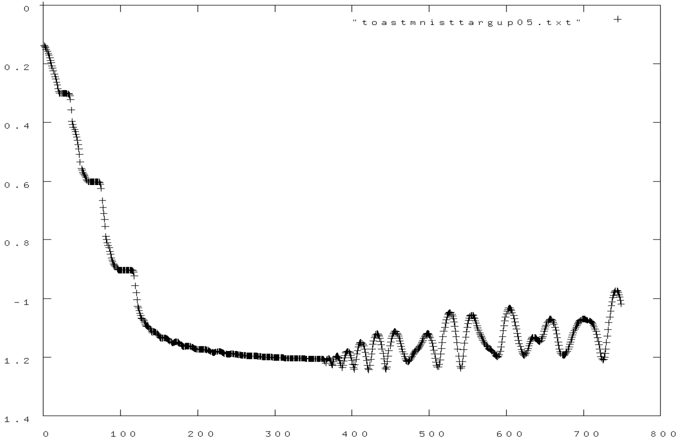


Figure 2: Reached value 0.591576 whereas target level 0.519984 unreachable.

SAVVY on MNIST, $n = 784, m = 10, d = 60000$



Resulting accuracy of one layer model with smooth-max activation and cross entropy loss on test set of 10000 images is the "optimal" 92%.

Mixed Binary Linear Optimization

Consider a piecewise linear optimization problem in Abs-Linear-Form

$$\text{Min}\, a^\top x + b^\top z + c^\top \Sigma z \quad s.t. \quad z = Zx + Mz + L\Sigma z \quad \text{and} \quad \Sigma z \geq 0$$

where $\sigma \in \{-1, 1\}^n$ and $\Sigma = \mathrm{diag}(\sigma)$ are binary variables.
This (**MIBLOP**) can be (**MILOP**), provided $|z|_\infty \leq \gamma$ yielding

$$\min_{x,z,w,\sigma} \quad \left(a^\top x + b^\top z + c^\top h + \tfrac{q}{2}\|x\|^2\right) \quad s.t. \quad z = Zx + Mz + Lh\,, \tag{2}$$
$$-h \leq z \leq h \quad \text{and} \quad h + \gamma(\sigma - e) \leq z \leq -h + \gamma(\sigma + e),$$

Quote by Fischetti and Jo (2018)

"Deep Neural Networks as 0-1 Mixed Integer Linear Programs: A Feasibility Study": PL models are unfortunately not suited for training.

Prediction by PL functions in ANF

For $x \in \mathbb{R}^n \mapsto y \in \mathbb{R}^m$

Continuous PL function $\iff$ Hinged NN $\iff$ Abs-Linear-Form .

Numb. of Layers $\ell \geq \nu =$ Switching Depth $=$ nilpotency of $(I - M)^{-1}L$.

$$\begin{aligned} z &= c + Zx + Mz + L|z| \in \mathbb{R}^s \\ y &= b + Jx + Nz \in \mathbb{R}^m \end{aligned}$$

1. where $M, L \in \mathbb{R}^{s \times s}$ are strictly lower triangular to yield $z = z(x)$.
2. $\equiv$ NN if $M \equiv L$ are block bidiagonal, other sparsity patterns possible.
3. note that $max(u, w) = u + (z + |z|)/2$ with $z = (w - u)$.
4. ALFs with $\nu \leq \bar{\nu}$ form infinite dimensional linear space of $C^{0,1}(\mathbb{R}^n)$.
5. ALFs can be successively abs-linearized with respect to $w = [c, Z, M, L, b, J, N]$ for learning=fitting.

Structured Piecewise linearization (PL) w.r.t. weight vector

Given a reference point $\mathring{w} = [\mathring{c}, \mathring{Z}, \mathring{M}, \mathring{L}, \mathring{b}, \mathring{J}, \mathring{N}]$ we have Taylor like

$$\tilde{z} = \mathring{z} + \Delta z(\mathring{w}; w - \mathring{w}) \quad \text{for} \quad x \text{ fixed}$$

where $\tilde{z}$ can be calculated directly from Abs-Linear-Form

$$\tilde{z} = [c + Zx + \Delta M \mathring{z} + \Delta L |\mathring{z}|] + \mathring{M}\tilde{z} + \mathring{L}|\tilde{z}|$$

with $\Delta M = M - \mathring{M}$, $\Delta L = L - \mathring{L}$. The discrepancy is bounded by

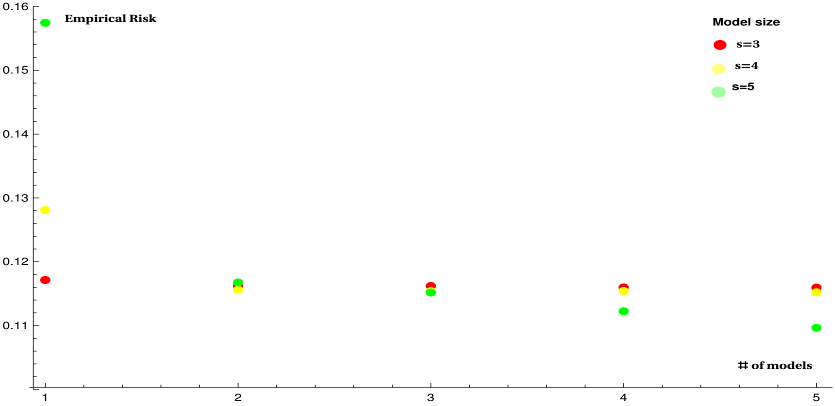
$$\|\tilde{z} - z\|_\infty \leq \tfrac{q}{2} \|\Delta M, \Delta L\|_F^2 .$$

Explicit upper bound on q can be given but seems too conservative.

Reverse Mode AD $\equiv$ Back Propagation yields at $* = 2\ \mathrm{OPS}(\tilde{z})$

$$\left[\bar{c}, \bar{Z}, \bar{M}, \bar{L}, \bar{b}, \bar{J}, \bar{N}\right] \equiv \frac{\partial(\bar{z}^\top \tilde{z})}{\partial[c, Z, \Delta M, \Delta L, b, J, N]} \, .$$

Objective for successive linearizations for model sizes 3,4,5,



Simplex Iterations by Gurobi

Regression on Griewank in 2 dimensions, 50 training data, 8 testing data over 5 successive piecewise linearizations.

s	#w	var.	1	2	3	4	5
3	21	471	303810	353703	1716277	581060	681025
4	31	631	1129639	263007	1015447	1339147	1068608
5	43	793	1153345	22793377	22895320	21241422	16513124

For s=5 there were 250 equality and 1000 inequality constraints, both linear.

Conclusion:

Nice try – but !!!

Question:

Are we overlooking any structure that could/should be exploited?

Potential contributions

1. SALMIN generates cluster points that are first oder minimal.
2. Analytically savvy ball reaches target level in convex case.
3. Savvy ball can climb away from undesirable local minimizers.
4. Successive PL allows exact integration of Savvy Ball and application of Mixed Binary Linear Optimization (Gurobi).
5. Though costly MIBLOP may provide reference solutions.
6. Stepsize chosen automatically via kinks and angle bound.
7. Abs-Linear-Learning generalizes hinged Neural Nets.

Improvements and Developments

1. Refine targeting and restarting strategy for SB.
2. Matrix based implementation for HPC with GPU.
3. Exploitation of low-rank updates in polyhedral transition.
4. Mini-batch version in stochastic gradient fashion.
5. Check global optimality of MIBLOP cluster points.
6. Piecewise linearize "loss"-function (e.g. sparsemax).
7. Adaptively enforce sparsity in Abs-Linear-Learning.

Muchas Gracias por su Atención !!