Generalized Abs-Linear Learning by Mixed Binary Quadratic Optimization

Andreas Griewank* — Ángel Rojas**

* Departement of Mathematics
Humboldt-Universität zu Berlin
Berlin
Germany
griewank@math.hu-berlin.de

** School of Mathematical and Computational Sciences
Yachay Tech
Urcuquí, Imbabura
Ecuador
angel.rojas@yachaytech.edu.ec

ABSTRACT. We consider predictor functions in generalized abs-linear form, which generalize neural nets with hinge activation. To train them with respect to a given data set of feature-label pairs, one has to minimize the average loss, which is a multi-piecewise linear or quadratic function of the weights, i.e. coefficients of the abs-linear form. We suggest to attack this nonsmooth, global optimization problem via successive piecewise linearization, which allows the application of mixed binary convex quadratic optimization codes amongst other methods. These solve the sequence of abs-linear model problems with a proximal term. Preliminary experiments on a simple regression problem verify the validity of the approach but require a large number of Simplex pivots by the solver Gurobi.

RÉSUMÉ. Nous considérons les fonctions prédictives sous une forme abs-linéaire généralisée, qui généralisent les réseaux neuronaux avec activation de la charnière. Pour les entraîner par rapport à un ensemble de données donné de paires étiquette-caractéristique, il faut minimiser la perte moyenne, qui est une fonction linéaire ou quadratique multi-morceaux des poids, c'est-à-dire des coefficients de la forme abs-linéaire. Nous suggérons d'attaquer ce problème d'optimisation globale non lisse via une linéarisation successive par morceaux, qui permet l'application de codes d'optimisation quadratiques convexes binaires mixtes entre autres méthodes. Ceux-ci résolvent la séquence de problèmes du modèle abs-linéaire avec un terme proximal. Des expériences préliminaires sur un problème de régression simple vérifient la validité de l'approche mais nécessitent un grand nombre de pivots Simplex par le solveur Gurobi

KEYWORDS : Abs-Normal/Linear Form, Successive Piecewise Linearization, Mixed Binary Linear/Quadratic Optimization, Proximal Term.

MOTS-CLÉS : Forme Abs-Normale / Linéaire, Linéarisation successive par morceaux, Optimisation binaire mixte linéaire / quadratique, Terme proximal.

1. Introduction and Notation

Neural nets with hinge activation have been proven theoretically [5] and experimentally [7] to produce prediction functions $f(w;x)$ that are able to represent a wide variety of relations in machine learning. These predictor models are piecewise linear [9] with respect to the feature vector x and multi-piecewise-linear (see explanation below) with respect to the weight vector w, which consists of various transformation matrices and inhomogeneous shifts. It is well known that every piecewise linear vector function from $x \in \mathbb{R}^n$ to $y \in R^m$ can be expressed in an abs-linear form

$$y = f(w;x) \equiv b + Jx + Nz + Y|z| \quad s.t. \quad z = F(w;x) \;=\; c + Zx + Mz + L|z| \quad (1)$$

where $z \in \mathbb{R}^s$ is a vector of switching variables and the various coefficient vectors and matrices can be combined to the *weight* vector

$$w \;\equiv\; (c, Z, M, L, b, J, N, Y) \in \mathbb{R}^{(s,s\times n,s\times s,s\times s,m,m\times n,m\times s,m\times s)} \;\simeq\; \mathbb{R}^{\bar{s}}$$

with $\bar{s} = (s+m)(1+n+2s)$. To make sure that $f(w;x)$ can be unambiguously evaluated as a piecewise linear continuous function of x we assume throughout that the square matrices $M, L \in \mathbb{R}^{s\times s}$ are strictly lower triangular.

In the case of multi-layer neural networks z represents nodal values and the matrices M and L are block diagonal, with $M = L$ in the case of hinge activation. By induction on the s components z_i of z one can easily see that they are piecewise linear functions $z_i(x)$ of x, which then also holds for the resulting vector

$$y = f(w;x) = b + Jx + Nz(x) + Y|z(x)| \in \mathbb{R}^m \,.$$

However, it is important to note that the dependance of $f(w;x)$ on the coefficients w is only *multi-piecewise linear*, i.e. piecewise linear with respect to each component of w when the others are kept constant. In other words for each Cartesian basis vector $e_j \in \mathbb{R}^{\bar{s}}$ and fixed x the univariate function $f(w + te_j; x)$ is piecewise linear. Therefore one can rather efficiently perform global coordinate searches even when the loss function on y is not piecewise linear but for example quadratic.

2. Separation, Scaling and Fixed Point Iteration

For the subsequent optimization approaches we normalize the abs-linear system by relegating the nonsmoothness to the second part of (1) and then scaling the z so that we can compute Lipschitz constants and an upper bound for $z(x)$.

Separation of Nonsmoothness:

The smoothness condition $Y = 0$ can always be achieved by extending the originally given switching vector z to $\tilde{z}^\top = (z^\top, |z|^\top Y^\top)$ and correspondingly extending L, M and N to

$$\tilde{L} = \begin{bmatrix} L & 0 \\ Y & 0 \end{bmatrix} \in \mathbb{R}^{(s+m)\times(s+m)} \ni \tilde{M} = \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \text{ and } \quad \tilde{N} = \begin{bmatrix} N & I \end{bmatrix} \in \mathbb{R}^{m\times(s+n)} \,.$$

Naturally, we also have to pad the vector $c \in \mathbb{R}^s$ and the matrix $Z \in \mathbb{R}^{s\times n}$ by m additional zero rows. Now the dimension of the coefficient vector $w \equiv (c, Z, M, L, b, J, N)$ has gone down by ms to $(s+m)(1+n)+s(m+2s)$ after the incrementation of s of course. Thus a typical empirical risk function may look like

$$\varphi(w) \equiv \tfrac{1}{2\bar{k}} \sum_{k=1}^{\bar{k}} \|f(w, x_k) - \tilde{y}_k\|_2^2 \quad \text{with} \quad f(w;x) = Nz(w;x) \tag{2}$$

where the (feature, label) pairs $(x_k, \tilde{y}_k) \in \mathbb{R}^{n\times m}$ for $k = 1, 2, \ldots \bar{k}$ form a suitable training set. Here N is typically a projection onto the last components of z.

Scaling for Contraction:

Secondly we can scale z to $\tilde{z} = Dz$ with $D = \mathrm{diag}(d)$ a matrix of positive entries $0 < d_i$ for $i = 1 \ldots s$. Then the state equation can be rewritten as

$$Dz \equiv \tilde{z} = \tilde{c} + \tilde{Z}x + \tilde{M}\tilde{z} + \tilde{L}|\tilde{z}| \equiv Dc + DZx + (DMD^{-1})Dz + (DLD^{-1})|Dz| \,.$$

Hence we see that the strictly lower triangular matrices M and L undergo a positive diagonal similarity transformation. Starting from $d_i = 1$ the d_i can be chosen for $i = 2 \ldots s$ such that the rows of $\tilde{M}$ and $\tilde{L}$ have ℓ_1 norms less than or equal to any given bound $\frac{\varepsilon}{2}$ with $\varepsilon < 1$. This implies for their ℓ_∞ matrix norms

$$\varepsilon \;\geq\; \|\tilde{M}\|_\infty + \|\tilde{L}\|_\infty \;\geq\; \||\tilde{M}| + |\tilde{L}|\|_\infty \;=\; \|\tilde{M}, \tilde{L}\|_\infty$$

Obviously, it is sufficient that the sums of the ℓ_1 norms of corresponding rows in $\tilde{M}$ and $\tilde{L}$, which equal to the rows obtained by summing or concatenating $|M|$ and $|L|$ are less than or equal to ε. Note that if scaling by D is performed, then the matrix N in the response must be transformed to ND^{-1}. Assuming without loss of generality that the norm bounds already hold for the original M and L we find that the map

$$G(z) \equiv G_{(M,L)}(z) \equiv z - H(z) \quad \text{with} \quad H(z) = Mz + L|z| \tag{3}$$

is a perturbation of the identity by the map $H(z)$. The latter has the componentwise Lipschitz property

$$\begin{aligned} |H(z) - H(\mathring{z})| &= |M(z-\mathring{z}) + L(|z| - |\mathring{z}|)| \leq |M||z-\mathring{z}| + |L||\,|z| - |\mathring{z}|\,| \\ &\leq |M||z-\mathring{z}| + |L||z-\mathring{z}| = (|M| + |L|)|z-\mathring{z}| \in \mathbb{R}^s \end{aligned} \tag{4}$$

where we have used that by the inverse triangle inequality $||z| - |\mathring{z}|| \leq |z - \mathring{z}|$ componentwise. This implies for the infinity norms of matrices and vectors

$$\|H(z) - H(\mathring{z})\|_\infty \;\leq\; \|M, L\|_\infty \|z - \mathring{z}\|_\infty \;\leq\; \varepsilon \|z - \mathring{z}\|_\infty \in \mathbb{R}$$

so that we have a contraction provided $\varepsilon < 1$. Thus it follows as a consequence of the Banach fixed point theorem that there is a unique point $z^* = c + Zx + H(z^*)$, and G has an inverse G^{-1} with the Lipschitz constant $1/(1-\varepsilon) < \infty$. Hence we obtain for the exact solution z^* the bound

$$z^* \;=\; G^{-1}_{(M,L)}(c + Zx) \;\Longrightarrow\; \|z^*\|_\infty \leq \|G^{-1}_{(M,L)}(c)\|_\infty + \frac{\|Z\|_\infty \|x\|_\infty}{(1 - \|M, L\|_\infty)} \,. \tag{5}$$

Fixed Point Iteration:

Moreover starting from z^0 we obtain for the iteration

$$z^{k+1} = c + Zx + Mz^k + L|z^k| \;=\; c + Zx + H(z^k) \tag{6}$$

from (4) the componentwise bound

$$|z^k - z^*| \;\leq\; (|M| + |L|)|z^{k-1} - z^*| \;\leq\; (|M| + |L|)^k |z^0 - z^*| \,. \tag{7}$$

In fact we do not only have a contraction but can establish finite convergence as follows. Since $|M| + |L|$ is strictly lower triangular we know that there is a minimal integer ν such that

$$(|M| + |L|)^\nu = 0 \quad \text{with} \quad \nu \leq s \quad \text{and} \quad \nu = 0 \iff M = 0 = L \,.$$

We call ν the switching depth since it counts how often the nonmooth elemental abs() can be super imposed on each other. In terms of the evaluation graph [4] it is the longest directed path along abs() nodes. If $\nu = 0$ the prediction function is smooth and thus linear, and if $\nu = 1$ we say that the problem is simply switched, which is typical for KKT like systems. In the case of neural networks ν is simply the number of intermediate layers.

Because of the bound (7), the iteration (6) must reach $z^k = z^*$ after at most ν steps from any z^0. This applies irrespective of the scaling. However, in general only the norm $\|D(z^k - z^*)\|_\infty$, and not the norm $\|z^k - z^*\|_\infty$, will decline monotonically. Of course from a good warm start we may hope to converge to a satisfactory accuracy in fewer than ν steps. We also emphasize that a matrix vector product can be much better computed in parallel rather than forward and backward substitutions on triangular matrices.

3. Mixed Binary Linear/Quadratic Optimization

For fixed w the constraints in (1) are essentially linear, except for the absolute values. Using general nonsmooth optimization methods in this situation would of course disregard a lot of structure. We will reformulate the problem as a mixed binary linear optimization problem, for which standard solvers are available.

Formulation w.r.t. $\mathbf{x} \in \mathbb{R}^\mathbf{n}$:

Assuming without loss of generality that $Y = 0$ as justified in Section 2 we can concentrate on the triangular *state equation*

$$z = c + Zx + Mz + L|z| \;=\; c + Zx + Mz + La \quad \text{with} \quad a = |z| \in \mathbb{R}^s \,. \tag{8}$$

In order to eliminate the highly nonlinear constraint $a = |z|$ we use a *signature matrix* $\Sigma = \mathrm{diag}(\sigma)$ whose elements $\sigma_i \in \{-1, 1\}$ for $i = 1 \ldots s$ are *binary variables*. Then the components of the vector $a = \Sigma z = (\sigma_i z_i)_{i=1\ldots s}$ are bilinear in the binary variable σ_i and the real variables z_i. It is well known [3] that the constraints $a = \Sigma z \geq 0$ can be linearized to the system of inequalties

$$-a \leq z \leq a \quad \text{and} \quad a + \gamma(\sigma - e) \leq z \leq -a + \gamma(\sigma + e) \,, \tag{9}$$

where $e \in \mathbb{R}^s$ is the vector of ones and $\gamma \in \mathbb{R}$ an upper bound on the possible norm values $\|z\|_\infty$. It can be computed by maximizing the right hand side of (5) over a range of x that are likely to occur.

From the property $\gamma \geq \|z\|_\infty$ it can be easily derived that these four vector inequalities are equivalent to the nonlinear constraint $a = |z|$. The key property of the inequalities (9) is that together with the equation $z = c + Zx + Mz + La$ it can be entered into Gurobi and other modern mixed integer optimization solvers to characterize a feasible set with respect to the real variable vector $(x, z, a) \in \mathbb{R}^{n+2s}$ and the binary variable vector $\sigma \in \{-1, 1\}^s$.

The corresponding objective must be linear or convex quadratic with respect to the real variables so that for example

$$f(w; x) \;=\; b + Jx + Nz(x) + \tfrac{1}{2}(x - \mathring{x})^\top Q(x - \mathring{x}) \in \mathbb{R} \tag{10}$$

with some positive semi-definite matrix $Q \in \mathbb{R}^{n\times n}$ and a reference point $\mathring{x} \in \mathbb{R}^n$. We can also include quadratic terms with respect to z, which we will do later in the learning context. Frequently we may simply have an Euclidean proximal term in that $Q = qI$ for some scalar $0 \leq q \in \mathbb{R}$. Given any parameter $q > 0$ the objective will be bounded below and a single call to Gurobi will yield a global minimizer.

Proximal Term via Quadratic Overestimation

One possible origin of such a proximal term is that we have originally a nonlinear state equation

$$z = F(x, z, |z|) \quad \text{with} \quad F \in C^2(\mathbb{R}^{(n+2s)}, \mathbb{R}^s) \quad . \tag{11}$$

It can be abs-linearized at a reference point $\mathring{x}$ with $\mathring{z} = z(\mathring{x})$ to

$$\tilde{z} = \mathring{z} + \mathring{Z}(x - \mathring{x}) + \mathring{M}(\tilde{z} - \mathring{z}) + \mathring{L}(|\tilde{z}| - |\mathring{z}|) \tag{12}$$

where the matrices

$$\mathring{Z} = \frac{\partial}{\partial x} F(x, z, |z|) \in \mathbb{R}^{(s\times n)},\; \mathring{M} = \frac{\partial}{\partial z} F(x, z, |z|) \in \mathbb{R}^{(s\times s)} \ni \mathring{L} = \frac{\partial}{\partial |z|} F(x, z, |z|)$$

are all evaluated at the point $(\mathring{x}, \mathring{z}, |\mathring{z}|)$. Again the square matrices $\mathring{M}, \mathring{L}$ of order s must be structurally strictly lower triangular so that $z(x)$ and $\tilde{z}(x)$ can be evaluated unambiguously by (11) and (12), respectively. As shown in [2] the error in this abs-linearization can be bounded by

$$\|\tilde{z}(x) - z(x)\|_\infty \;\leq\; \tfrac{q}{2}\|x - \mathring{x}\|^2 \quad \text{for some } q > 0 \; .$$

If the original scalar response was just $f(w; x) = b + Jx + Nz(x)$ we get the upper bound

$$\begin{aligned} f(w; x) \quad &\leq \quad b + Jx + N\tilde{z}(x) + \|N\|_\infty \|z(x) - \tilde{z}(x)\| \\ \leq \tilde{f}(w; x) \quad &\equiv \quad b + Jx + N\tilde{z}(x) + \tfrac{\tilde{q}}{2}\|x - \mathring{x}\|^2 \quad \text{with} \quad \tilde{q} = q\|N\|_\infty \; . \end{aligned} \tag{13}$$

Hence we have exactly an objective of the proxlinear form (10) with $\tilde{z}(x)$ defined by the abs-linear triangular system (12). It can of course be rewritten as a mixed binary linear system using the reformulation (9) and then solved by Gurobi.

While upper bounds for $\tilde{q} = q\|N\|_\infty$ can be derived from the evaluation procedure of F at $(\mathring{x}, \mathring{z}, |\mathring{z}|)$ it may also be updated iteratively like the size parameters in trust region or quadratic overestimation methods. As usual being conservative, i.e. making $\tilde{q}$ large enforces convergence but may slow down the actual computation.

Outer Loop:

So the full method consists of forming the constraints (12) at the current iterate $\mathring{x}$, replacing $|\tilde{z}|$ with an a satisfying (9) and minimizing the objective (13) by Mixed Binary Quadratic Optimization. The global minimizer $\hat{x}$ can then serve as the new reference point $\mathring{x}$ provided we have for the real objective

$$b + J\hat{x} + Nz(\hat{x}) < b + J\mathring{x} + Nz(\mathring{x}) \ .$$

Otherwise the estimate for $\tilde{q}$ was too small and must be increased significantly without a change in $\mathring{x}$. As shown in [8] for a method that only computes local minimizers of the successive abs-linearizations, all clusterpoints x_* of the sequence generated in this way are first order minimal, i.e. minimizers of the abs-linearization at x_*. Unfortunately, even globally solving each local abs-linear problem does in general not guarantee that the cluster point is a global minimizer of the underlying problem.

Piecewise Linearization w.r.t. w:

So far we have assumed that the coefficients $w = (c, Z, M, L)$ are constant throughout the minimization. For abs-linear learning we cannot make this assumption and rather have to consider w as variable and x as constant. Now the vector constraint $z = c + Zx + Mz + La$ contains the terms Mz and La which are bilinear in the components of the real variables (z, a) and w. That means unfortunately that they cannot be linearized exactly. To overcome this difficulty we perform a piecewise linearization of the original state equation $z = c + Zx + Mz + L|z|$ at a reference point $(\mathring{c}, \mathring{Z}, \mathring{M}, \mathring{L}, \mathring{z})$ with x fixed and $\mathring{z}$ the corresponding z value such that $\mathring{z} = \mathring{c} + \mathring{Z}x + \mathring{M}\mathring{z} + \mathring{L}|\mathring{z}|$. Thus we obtain with G as defined in (3) immediately the relation

$$G(\mathring{z}) = G_{(M,L)}(\mathring{z}) = \mathring{c} + \mathring{Z}x - \Delta M\mathring{z} - \Delta L|\mathring{z}|$$

where naturally $\Delta M = M - \mathring{M}$ and $\Delta L = L - \mathring{L}$. Assuming that $\mathring{M}$ and $\mathring{L}$ are scaled as described above and also $\|M, L\|_\infty \leq \varepsilon < 1$ we find that the inverse operator $G^{-1} = G^{-1}_{(M,L)}$ still has the Lipschitz constant $1/(1-\varepsilon)$.

Then we obtain for $\Delta z = z - \mathring{z}$ the bound

$$\begin{aligned} \|\Delta z\|_\infty &\leq \frac{\|\Delta c + \Delta Z x + \Delta M\mathring{z} + \Delta L|\mathring{z}|\|_\infty}{1 - \|M, L\|_\infty} \\ &\leq \frac{\|\Delta c\|_\infty + \|\Delta Z\|_\infty\|x\|_\infty + \|\Delta M, \Delta L\|_\infty\|\mathring{z}\|_\infty}{1 - \|M, L\|_\infty} \end{aligned} \tag{14}$$

where naturally $\Delta c = c - \mathring{c}$ and $\Delta Z = Z - \mathring{Z}$.

Now we do want to find a piecewise linearization of the abs-normal function

$$z = \hat{F}(w) = G^{-1}_{(M,L)}(c + Zx)$$

i.e. an abs-linear function $\tilde{z} = \mathring{F}(w)$ approximation in the variables $w = (c, Z, M, L)$.

Proposition. *The vector function*

$$\tilde{z} \;=\; \mathring{F}(w) \;\equiv\; G^{-1}_{(\mathring{M},\mathring{L})}(c + Zx + \Delta M\mathring{z} + \Delta L|\mathring{z}|)$$

is abs-linear and can be evaluated by solving the triangular system

$$\tilde{z} = (c + Zx + \Delta M\mathring{z} + \Delta L|\mathring{z}|) + \mathring{M}\tilde{z} + \mathring{L}|\tilde{z}| \;. \tag{15}$$

Moreover, provided

$$\|\mathring{M}, \mathring{L}\|_\infty + \|\Delta M, \Delta L\|_\infty \le \varepsilon < 1 \tag{16}$$

the discrepancy between $\hat{F}(w)$ and its abs-linearization $\mathring{F}(w)$ is bounded by

$$\|\hat{F}(w) - \mathring{F}(w)\|_\infty \;\le\; \frac{\|\mathring{z}\|_\infty}{(1-\varepsilon)^2}\|\Delta M, \Delta L\|^2_\infty \;\le\; \frac{s\|\mathring{z}\|_\infty}{(1-\varepsilon)^2}\|\Delta M, \Delta L\|^2_F \;.$$

The proof of this result can be found in the appendix Section 6.

The overall Mixed Binary Quadratic Problem:

Using $(z+\tilde{z})^\top N^\top N(z-\tilde{z}) = \|Nz\|^2 - \|N\tilde{z}\|^2$ we obtain for a single loss term in (2)

$$\begin{aligned} & \tfrac{1}{2}\|Nz-\tilde{y}\|_2^2 \;-\; \tfrac{1}{2}\|N\tilde{z}-\tilde{y}\|_2^2 \;=\; \tfrac{1}{2}(N(z+\tilde{z})-2\tilde{y})^\top N(z-\tilde{z}) \\ \le\; & \|N\mathring{z}-\tilde{y}\|_2\|N(z-\tilde{z})\|_2 + O(\|\Delta w\|_2^3) \\ \le\; & \|N\mathring{z}-\tilde{y}\|_2\|N\|_1\frac{s\|\mathring{z}\|_\infty}{(1-\varepsilon)^2}\|\Delta M, \Delta L\|^2_F + O(\|\Delta w\|_2^3) \;. \end{aligned} \tag{17}$$

For each one of the sample points $(x_k, \tilde{y}_k)$ we get a different $\tilde{z} = \tilde{z}(w; x_k)$ which will be denoted simply by $z_k(w)$. Then the resulting mixed binary quadratic problem has the convex quadratic objective

$$\varphi(w) \;\equiv\; \frac{1}{2\bar{k}}\sum_{k=1}^{\bar{k}}\|Nz_k(w) - \tilde{y}_k\|_2^2 + \frac{s\|N\|_1\|\Delta w\|_2^2}{\bar{k}(1-\varepsilon)^2}\sum_{k=1}^{\bar{k}}\|\mathring{z}_k\|_\infty\|N\mathring{z}_k - \tilde{y}_k\|_2 \;. \tag{18}$$

So the first term is the approximation to the original quadratic loss risk via the piecewise linearization of the state equation and the second term is a proximal quadratic that tries to bound the resulting error in the objective.

For each data index $k = 1 \ldots \bar{k}$ we get with $a_k \in \mathbb{R}^s$ and $\sigma_k \in \{-1, +1\}^s$ the bilinear constraint set

$$\begin{aligned} z_k \quad &= c + Zx + \Delta M\mathring{z}_k + \Delta L|\mathring{z}_k| + \mathring{M}z_k + \mathring{L}a_k \\ -a_k \quad &\le z_k \le a_k \quad \text{and} \quad a_k + \gamma(\sigma_k - e) \le z_k \le -a_k + \gamma(\sigma_k + e) \;. \end{aligned} \tag{19}$$

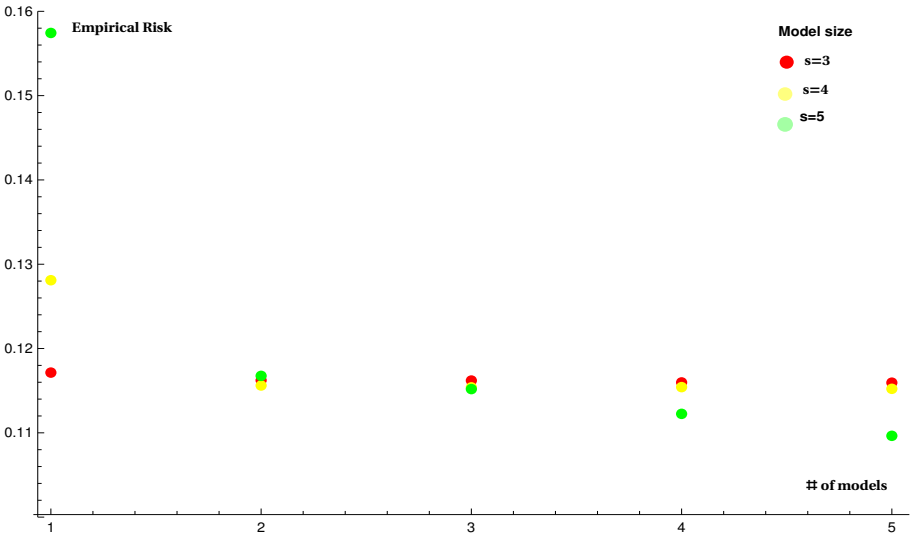
Hence we see that we have the same binary variables as in the formulation w.r.t. x and the only thing that changes compared to (8) is that we have the new linear terms $\Delta M\mathring{z}$ and $\Delta L|\mathring{z}|$. The bound γ can of course be adjusted individually to γ_k for each k. The minimizing of (18) subject to the linear constraints (19) and (9) with the σ_k binary can be delegated to solvers like Gurobi. The global minimizer $\hat{w}$ of the model problem can then serve as the reference point $\mathring{w}$ for the next abs-linearization, where we have to recompute the proximal term in (18).

4. Preliminary Numerical Experiment and Conclusion

Given the limitations of space and time we have not been able to conduct and report numerical results that are in any way conclusive. We simply minimized the averaged loss defined in (2) for the Griewank function [1]

$$\tilde{y}(x) = 1 + \frac{1}{4000}\sum_{k=1}^{d} x_k^2 - \prod_{k=1}^{d} \cos\left(\frac{x_k}{\sqrt{k}}\right)$$

over 50 training points and 8 testing points $x \in \mathbb{R}^2$ chosen uniformly at random in the cube $[-8, 8]^d$. The results reported are for the model sizes $s = 3, 4, 5$ over 5 successive linearizations. This results in 15 mixed integer quadratic optimization problems specified in (2) and (19). The following graph shows the true objective function values achieved by successive linearizations.



As one can see for the two smaller prediction models the minimal objective function is essentially already reached at the second linearization whereas the model with $s = 5$ switching variables can bring the objective further down. The number of weights, Gurobi variables and Simplex iterations are listed in the following table. The bad news is that the latter appears to be growing rapidly with respect to the model size. Nevertheless the exact solution can serve as a reference point for other methods and hopefully Gurobi can be accelerated by exploiting more of the special structure.

s	#w	#variables	pl 1	pl 2	pl 3	pl 4	pl 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

5. References

[1] A. GRIEWANK, "Generalized descent for global optimization", *Journal of Optimization Theory and Applications*, vol. 34, 1981.

[2] A. GRIEWANK, "On Stable Piecewise Linearization and Generalized Algorithmic Differentiation", *Optimization Methods and Software*, vol. 28, num. 6, 2013.

[3] A. GUPTE, S. AHMED, M. CHEON, AND S. DEY, "Solving Mixed Integer Bilinear Problems Using MILP Formulations", *SIAM Journal on Optimization*, vol. 23, num. 2, 2013.

[4] A. GRIEWANK AND A. WALTHER, "Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation", *SIAM*, 2nd Edition, 2008.

[5] D. YAROTSKY, "Error bounds for approximations with deep ReLU networks", *CoRR*, vol. abs/1610.01145, 2016.

[6] GRIEWANK, A., ROJAS, Á., "(2019) Treating Artificial Neural Net Training as a Nonsmooth Global Optimization Problem", *In: Nicosia G., Pardalos P., Umeton R., Giuffrida G., Sciacca V. (eds) Machine Learning, Optimization, and Data Science. LOD 2019. Lecture Notes in Computer Science*, vol. 11943. Springer, Cham.

[7] H. BÖLCSKEI, P. GROHS, G. KUTYNIOK, AND P. PETERSEN, "Optimal Approximation with Sparsely Connected Deep Neural Networks", *SIAM Journal on Mathematics of Data Science*, 2019.

[8] S. FIEGE, A. WALTHER AND A. GRIEWANK, "An Algorithm for Nonsmooth Optimization by Successive Piecewise Linearization", *Mathematical Programming*, 2018.

[9] S. SCHOLTES, "Introduction to Piecewise Differentiable Functions", *Springer*, 2012.

6. Appendix

Proof(Proposition). *According to* (8) *the exact solution* $z = \hat{F}(w)$ *satisfies the triangular system* $z - \mathring{M}z - \mathring{L}|z| = c + Zx + \Delta Mz + \Delta L|z|$ *and thus*

$$z \;=\; G^{-1}_{(\mathring{M},\mathring{L})}(c + Zx + \Delta Mz + \Delta L|z|)\;.$$

The difference between the two right hand sides of $\tilde{z}$ *and* z *is* $\Delta M(\mathring{z} - z) + \Delta L(|\mathring{z}| - |z|)$ *so that by the Lipschitz continuity of* $G^{-1}_{(\mathring{M},\mathring{L})}$

$$\|\tilde{z} - z\|_\infty \quad\leq\quad \frac{\|\Delta M(\mathring{z} - z) + \Delta L(|\mathring{z}| - |z|)\|_\infty}{(1 - \|\mathring{M}, \mathring{L}\|_\infty)} \;\leq\; \frac{\|\Delta M, \Delta L\|_\infty \|\Delta z\|_\infty}{(1 - \|\mathring{M}, \mathring{L}\|_\infty)}\;.$$

Since with $\Delta\tilde{z} = \tilde{z} - \mathring{z}$

$$\|\Delta z\|_\infty \leq \|\Delta\tilde{z}\|_\infty + \|\Delta z - \Delta\tilde{z}\|_\infty = \|\Delta\tilde{z}\|_\infty + \|\tilde{z} - z\|_\infty$$

we can collect the terms in $\|\tilde{z} - z\|_\infty$ *on the left hand side and then divide both sides by its factor yielding*

$$\|\tilde{z} - z\|_\infty \;\leq\; \frac{\|\Delta M, \Delta L\|_\infty \|\Delta\tilde{z}\|_\infty}{(1 - \|\mathring{M}, \mathring{L}\|_\infty - \|\Delta M, \Delta L\|_\infty)}\;.$$

The penultimate inequality follows since again by Lipschitz continuity of $G^{-1}_{(\mathring{M},\mathring{L})}$ *between the arguments* $c + Zx$ *and* $c + Zx + \Delta M\mathring{z} + \Delta L|\mathring{z}|$

$$\|\Delta\tilde{z}\|_\infty \;\leq\; \frac{\|\Delta M\mathring{z} + \Delta L|\mathring{z}|\|_\infty}{(1 - \|\mathring{M}, \mathring{L}\|_\infty)} \;\leq\; \frac{\|\Delta M, \Delta L\|_\infty \|\mathring{z}\|_\infty}{(1 - \|\mathring{M}, \mathring{L}\|_\infty)}\;.$$

For the final bound we use that the the ℓ_∞ *norm of a* $(2s \times s)$ *matrix is the maximal* ℓ_1 *norm of any one of its* s *rows, which is the bounded by* $\sqrt{2s}$ *times its* ℓ_2 *vector norm, which in turn is bounded by the Frobenius norm of the full matrix. This bound is sharp when* $\Delta M, \Delta L$ *has one nontrivial row of constants.*