An Inexact Augmented Lagrangian Algorithm for Training Leaky ReLU Neural Network with Group Sparsity

Wei Liu
liuw16@rpi.edu
Rensselaer Polytechnic Institute

Joint work with **Xin Liu** (AMSS, CAS), and **Xiaojun Chen** (PolyU)

Talk in SIAM OP23
June 2, 2023

References

Wei Liu, Xin Liu, and Xiaojun Chen, *Linearly Constrained Nonsmooth Optimization for Training Autoencoders.* SIAM Journal on Optimization, 2022, 32(3): 1931-1957

Wei Liu, Xin Liu, and Xiaojun Chen, *An Inexact Augmented Lagrangian algorithm for Training Leaky ReLU Neural Network with Group Sparsity.* Submitted to Journal of Machine Learning Research, under minor revision

1. The optimization problem for training deep neural networks

The Model for The Neural Network

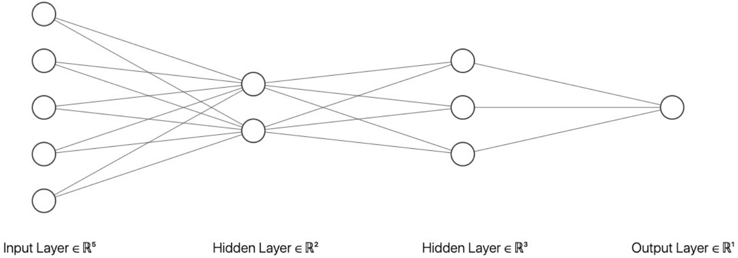
Neural Networks (NN)– deep learning

Given an input data $\{(x_n, y_n)\}_{n=1}^N$, where $x_n \in \mathbb{R}^{N_0}$, $y_n \in \mathbb{R}^{N_L}$

$$\min_{\substack{W_\ell, b_\ell, \\ \ell=1,2,\ldots,L}} \frac{1}{N} \sum_{n=1}^{N} \|\sigma(W_L\sigma(\cdots\sigma(W_1 x_n + b_1) + b_2 \cdots) + b_L) - y_n\|^2$$

- N: the number of input data
- L: the number of layers
- N_ℓ: the number of neurons at the ℓ-th layer
- σ: the activation function.
  Here we use leaky ReLU, i.e., $\sigma(z) = \max(z, \alpha z)$ with $1 > \alpha > 0$
- variables to be determined:
  - $W_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$
  - $b_\ell \in \mathbb{R}^{N_\ell}$

An Example



Why leaky ReLU?

- ReLU and leaky ReLU reduce the vanishing gradient phenomenon by making the hidden layers sparse ([Sun 2019])
- the performance of the leaky ReLU network is reported to be slightly better than that of the ReLU network ([Mass et al. 2013, Pedamonti 2018])

Our Aimed Problem

DNN with group sparsity

$$(1) \qquad \min_{w,b} \frac{1}{N} \sum_{n=1}^{N} \|\sigma(W_L \sigma(\cdots \sigma(W_1 x_n + b_1) + \cdots) + b_L) - y_n\|^2 + \mathcal{R}_1(w)$$

- $\mathcal{R}_1$: group lasso regularizer with $\lambda_w > 0$, i.e.,

$$\mathcal{R}_1(w) := \lambda_w \sum_{\ell=1}^{L} \|W_\ell\|_{2,1} = \lambda_w \sum_{\ell=1}^{L} \sum_{j=1}^{N_{\ell-1}} \|(W_\ell)_{\cdot,j}\|$$

- $w = \left(\text{vec}(W_1)^{\mathrm{T}}, \ldots, \text{vec}(W_L)^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^{\widetilde{N}}, \ b = \left(b_1^{\mathrm{T}}, \ldots, b_L^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^{\overline{N}}$
- $\widetilde{N} := \sum_{\ell=1}^{L} N_\ell N_{\ell-1}$, $\overline{N} := \sum_{\ell=1}^{L} N_\ell$

Why group sparsity?

- pursuing the parameter sparsity and theoretical improvement in efficiency ([Hoefler et al. 2021])
- requiring less training time ([Wen et al. 2016])

Existing Approaches: SGD-based Methods

Properties

- calculating the gradient via the chain rule
- in cases where nonsmooth activation functions are used, a subgradient in a neighborhood is often used

Some SGD-based Methods

- Vanilla SGD ([Cramir 1946])
- Adagrad ([Duchi-Hazan-Singer 2011])
- AdagradDecay ([Duchi-Hazan-Singer 2011])
- Adadelata ([Zeiler 2012])
- Adam ([Kingma-Ba-Adam 2014])
- Adamax ([Kingma-Ba-Adam 2014])

Existing Approaches: SGD-based Methods (Cont'd)

Limitations

- neglecting the exactness in calculating the subgradient of the objective function ⇒ convergence?

chain rule does not hold

$$\min_{z\in\mathbb{R}} f(z) = \frac{1}{2}\left(-(zx_1)_+ + (zx_2)_+ + 1\right)^2 + \frac{1}{2}\left(-(zy_1)_+ + (zy_2)_+ + 1\right)^2$$

where $x = (-1, 1), y = (-2, 0), z_+ := \max\{0, z\}$.
SGD calculate

$$\begin{aligned} g(z) := &\left(-(zx_1)_+ + (zx_2)_+ + 1\right)\left(-x_1 h(zx_1) + x_2 h(zx_2)\right) \\ &+ \left(-(zy_1)_+ + (zy_2)_+ + 1\right)\left(-y_1 h(zy_1) + x_2 h(zy_2)\right) \end{aligned}$$

as one derivative of f at z, $h(z) := \mathrm{sign}(z_+)$.
however, 0 is not a stationary point!

More examples: [April-July-Kummer 2011]

An equivalent Model

[Carreira Perpiñán-Wang 2012]

$$
\begin{aligned}
(2) \qquad & \min_{\substack{W_\ell, v_{n,\ell}, b_\ell, \\ \ell=1,2,\ldots,L,\ n=1,2,\ldots,N}} && \frac{1}{N}\sum_{n=1}^{N} \|v_{n,L} - y_n\|_F^2 + \mathcal{R}_1(w) \\
& \text{s. t.} && v_{n,\ell} = \sigma_\ell(W_\ell v_{n,\ell-1} + b_\ell) \\
& && n = 1, \ldots, N,\ \ell = 1, \ldots, L
\end{aligned}
$$

- $v_{n,0} = x_n$
- $v_{n,\ell}$: the output of the ℓ-th layer with respect to the n-th input data

Solving (2) instead of (1) is able to alleviate the problems that SGD-based methods suffer

Existing Methods for Solving the Problem (2)

- ℓ_2 penalty method:
  - MAC [Carreira-Perpinan-Wang 2014]
  - proximal BCD (pBCD) [Lau-Zeng-Wu-Yao 2018]
  - pBCD [Zeng-Lau-Lin-Yao 2019]
- multi-block 'ADMM':
  - Taylor et. al. 2016
  - Zhang-Chen-Saligrama 2016
  - Evens-Latafat-Themelis 2020

The above methods enjoy no exact penalty results!

ℓ_1 penalty method: present the exact penalty result, and establish the subsequence convergence result ([Cui-He-Pang 2020])

- limitation in theory: excludes ReLU and leaky ReLU
- limitation in practice: no regularizer

$\Rightarrow$ Design some new ℓ_1 penalty method.

A New Formulation

$$
(P) \quad \begin{aligned} \min_{w,b,v,u} \ & \bar{O}(w,v) := \frac{1}{N}\sum_{n=1}^{N} \|v_{n,L} - y_n\|^2 + \mathcal{R}_1(w) + \mathcal{R}_2(v) \\ \text{s. t.} \ & \sigma(u_{n,\ell}) - v_{n,\ell} = 0,\ u_{n,\ell} - (W_\ell v_{n,\ell-1} + b_\ell) = 0, \\ & n \in [N],\ \ell \in [L]. \end{aligned}
$$

- $v := \left(v_{1,1}^{\mathrm{T}}, v_{2,1}^{\mathrm{T}}, \ldots, v_{1,L}^{\mathrm{T}}, v_{2,L}^{\mathrm{T}}, \ldots, v_{N,L}^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^m$
- $u = \left(u_{1,1}^{\mathrm{T}}, u_{2,1}^{\mathrm{T}}, \ldots, u_{1,L}^{\mathrm{T}}, u_{2,L}^{\mathrm{T}}, \ldots, u_{N,L}^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^m$
- $\mathcal{R}_2(v) = \lambda_v \|v\|^2$: a regularization term
- Feasible set: $\Omega_1 := \{(w,b,v,u) : v - \sigma(u) = 0, u = \Psi(v)w + Ab\}$
- linear operator $\Psi(v) : \mathbb{R}^m \mapsto \mathbb{R}^{m\times\widetilde{N}}$, matrix $A \in \mathbb{R}^{m\times\overline{N}}$
- $m = N\overline{N}$

A New Penalty Approach

- we consider to have $v \geq \sigma(u)$ as a constraint and add a penalty term $\beta^{\mathrm{T}}(v - \sigma(u))$ in the objective function
- $\beta = (\beta_1 e_{NN_1}^{\mathrm{T}}, \ldots, \beta_L e_{NN_L}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^m$
- we write $v \geq \sigma(u)$ by $v - u \geq 0$ and $v - \alpha u \geq 0$

$$(PP) \qquad \begin{aligned} \min_{w,b,v,u} \; & O(w,v,u) = \bar{O}(w,v) + \beta^{\mathrm{T}}(v - \sigma(u)) \\ \text{s. t.} \; & v - u \geq 0, v - \alpha u \geq 0, u = \Psi(v)w + Ab. \end{aligned}$$

- feasible set: Ω_2
- we focus on obtaining a l(imiting)-stationary point of problem (PP)

Definitions

- Clarke subdifferential: $\partial^c f(\bar{z}) = \mathrm{co}\{\lim_{z \to \bar{z}} \nabla f(z) : f \text{ is smooth at } z\}$
- limiting subdifferential: $\partial f(\bar{z}) :=$ $\left\{ v : \exists z^k \xrightarrow{f} \bar{z}, v^k \to v \text{ such that } \liminf_{z \to z^k} \frac{f(z) - f(z^k) - \langle v^k, z - z^k \rangle}{\|z - z^k\|} \geq 0, \ \forall k \right\}$,
- a point $\bar{z} \in \mathcal{Z}$ is a l-stationary point, a C(larke)-stationary point of $\min_{z \in \mathcal{Z}} f(z)$ if $0 \in \partial f(\bar{z}) + \mathcal{N}_{\mathcal{Z}}(\bar{z})$, $0 \in \partial^c f(\bar{z}) + \mathcal{N}_{\mathcal{Z}}^c(\bar{z})$, respectively

A New Penalty Approach

- we consider to have $v \geq \sigma(u)$ as a constraint and add a penalty term $\beta^{\mathrm{T}}(v - \sigma(u))$ in the objective function
- $\beta = (\beta_1 e_{NN_1}^{\mathrm{T}}, \ldots, \beta_L e_{NN_L}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^m$
- we write $v \geq \sigma(u)$ by $v - u \geq 0$ and $v - \alpha u \geq 0$

$$(PP) \qquad \begin{aligned} \min_{w,b,v,u} \ & O(w, v, u) = \bar{O}(w, v) + \beta^{\mathrm{T}}(v - \sigma(u)) \\ \text{s. t.} \ & v - u \geq 0, v - \alpha u \geq 0, u = \Psi(v)w + Ab. \end{aligned}$$

- feasible set: Ω_2
- we focus on obtaining a l(imiting)-stationary point of problem (PP)

Definitions

- Clarke subdifferential: $\partial^c f(\bar{z}) = \mathrm{co}\,\{\lim_{z \to \bar{z}} \nabla f(z) : f$ is smooth at $z\}$
- limiting subdifferential: $\partial f(\bar{z}) :=$ $\left\{ v : \exists z^k \xrightarrow{f} \bar{z}, v^k \to v \text{ such that } \liminf_{z \to z^k} \frac{f(z) - f(z^k) - \langle v^k, z - z^k \rangle}{\|z - z^k\|} \geq 0, \ \forall k \right\}$,
- a point $\bar{z} \in \mathcal{Z}$ is a l-stationary point, a C(larke)-stationary point of $\min_{z \in \mathcal{Z}} f(z)$ if $0 \in \partial f(\bar{z}) + \mathcal{N}_{\mathcal{Z}}(\bar{z})$, $0 \in \partial^c f(\bar{z}) + \mathcal{N}_{\mathcal{Z}}^c(\bar{z})$, respectively

Why l-stationary point?

- l-stationary is stronger than C-stationary

An example

Consider
$$\min_{w_1\in\mathbb{R},w_2\in\mathbb{R},b_1\in\mathbb{R},b_2\in\mathbb{R}} f(w_1,w_2,b_1,b_2) := \\ ((w_2\sigma(w_1+b_1)+b_2)+1)^2 + ((w_2\sigma(2w_1+b_1)+b_2)-1)^2 . \tag{3}$$

let $w_2^*=1,\ b_1^*=0,\ w_1^*=0,\ b_2^*=0,$

$\partial^c f(w_1^*,w_2^*,b_1^*,b_2^*) = \left\{(t,0,s,0)^{\mathrm{T}} : t\in[2\alpha-4,2-4\alpha], s\in[-2+2\alpha,2-2\alpha]\right\},$

$\partial\left(f(w_1^*,w_2^*,b_1^*,b_2^*)\right)$

$= \left\{(-2\alpha,0,0,0)^{\mathrm{T}},(2\alpha-4,0,2\alpha-2,0)^{\mathrm{T}},(2-4\alpha,0,2-2\alpha,0)^{\mathrm{T}},(-2,0,0,0)^{\mathrm{T}}\right\},$

$f(w_1^*+\epsilon,w_2^*,b_1^*,b_2^*) = 5\epsilon^2-2\epsilon+2 < 2 = f(w_1^*,w_2^*,b_1^*,b_2^*),\ \epsilon:$ a small positive scalar

for any $0<\alpha<\frac{1}{2}$, $(w_1^*,w_2^*,b_1^*,b_2^*)$ is a C-stationary point, but is not a l-stationary point and local minimizer.

2. Theoretical Results

Stationary Point of (P)

$v - \sigma(u) = 0$ can be rewritten as complementary constraints:
$v - u \geq 0, (v - u)(v - \alpha u) = 0, v - \alpha u \geq 0$, We call $(w^*, b^*, v^*, u^*) \in \Omega_1$ a MPCC W-stationary point [Scheel-Scholtes 2000; Guo-Chen 2021] of (P), if there exist $\mu^1 \in \mathbb{R}^m$, $\mu^2 \in \mathbb{R}^m$ and $\xi \in \mathbb{R}^m$ such that

$$
\begin{aligned}
&0 = \nabla_w \bar{O}(w^*, v^*) + \Psi(v^*)^{\mathrm{T}} \xi, \quad 0 = A^{\mathrm{T}} \xi \\
&0 = \nabla_v \bar{O}(w^*, v^*) - \mu^1 - \mu^2 + \nabla_v \xi^{\mathrm{T}} (u^* - \Psi(v^*) w^*) \\
&0 = \mu^1 + \alpha \mu^2 + \xi \\
&(\mu^1)^{\mathrm{T}} (v^* - u^*) = 0, \ (\mu^2)^{\mathrm{T}} (v^* - \alpha u^*) = 0
\end{aligned}
$$

MPCC W-stationary point + $\mu_i^1 \mu_i^2 \geq 0, \forall i : u_i^* = v_i^* = 0 \Rightarrow$ MPCC C-stationary point

Lemma 1

NNAMCQ[a] holds for the constraints set of problem (P) $\Rightarrow$ Any local minimizer of (P) is its MPCC C-stationary point

[a][Ye-Zhang 2013]

Stationary Point of (P)

$v - \sigma(u) = 0$ can be rewritten as complementary constraints: $v - u \geq 0, (v-u)(v-\alpha u) = 0, v - \alpha u \geq 0$, We call $(w^*, b^*, v^*, u^*) \in \Omega_1$ a MPCC W-stationary point [Scheel-Scholtes 2000; Guo-Chen 2021] of (P), if there exist $\mu^1 \in \mathbb{R}^m$, $\mu^2 \in \mathbb{R}^m$ and $\xi \in \mathbb{R}^m$ such that

$$0 = \nabla_w \bar{O}(w^*, v^*) + \Psi(v^*)^{\mathrm{T}}\xi, \quad 0 = A^{\mathrm{T}}\xi$$

$$0 = \nabla_v \bar{O}(w^*, v^*) - \mu^1 - \mu^2 + \nabla_v \xi^{\mathrm{T}}(u^* - \Psi(v^*)w^*)$$

$$0 = \mu^1 + \alpha\mu^2 + \xi$$

$$(\mu^1)^{\mathrm{T}}(v^* - u^*) = 0, \ (\mu^2)^{\mathrm{T}}(v^* - \alpha u^*) = 0$$

MPCC W-stationary point + $\mu_i^1\mu_i^2 \geq 0, \forall i : u_i^* = v_i^* = 0 \Rightarrow$ MPCC C-stationary point

Lemma 1

NNAMCQ[a] holds for the constraints set of problem (P) $\Rightarrow$ Any local minimizer of (P) is its MPCC C-stationary point

[a] [Ye-Zhang 2013]

Stationary Point of (PP)

We call (w^*, b^*, v^*, u^*) a KKT point of (PP), if there exists $\mu^1 \in \mathbb{R}_+^m$, $\mu^2 \in \mathbb{R}_+^m$ and $\xi \in \mathbb{R}^m$ such that

$$
\begin{aligned}
&0 = \nabla_w \bar{O}(w^*, v^*) + \Psi(v^*)^{\mathrm{T}}\xi, \quad 0 = A^{\mathrm{T}}\xi \\
&0 = \nabla_v \bar{O}(w^*, v^*) + \beta - \mu^1 - \mu^2 + \nabla_v \xi^{\mathrm{T}}(u^* - \Psi(v^*)w^*) \\
&0 \in \partial_u(-\beta^{\mathrm{T}}\sigma(u^*)) + \mu^1 + \alpha\mu^2 + \xi \\
&(\mu^1)^{\mathrm{T}}(v^* - u^*) = 0, \ (\mu^2)^{\mathrm{T}}(v^* - \alpha u^*) = 0
\end{aligned}
$$

Lemma 2

MFCQ [a] holds for the constraints set of problem (PP).

[a] [Mangasarian 1994]

- $\Rightarrow$ (w^*, b^*, v^*, u^*) is a l-stationary point of (PP) if and only if (w^*, b^*, v^*, u^*) is a KKT point of (PP)
- $\Rightarrow$ any local minimizer of (PP) is its l-stationary point

Stationary Point of (PP)

We call (w^*, b^*, v^*, u^*) a KKT point of (PP), if there exists $\mu^1 \in \mathbb{R}^m_+$, $\mu^2 \in \mathbb{R}^m_+$ and $\xi \in \mathbb{R}^m$ such that

$$
\begin{aligned}
&0 = \nabla_w \bar{O}(w^*, v^*) + \Psi(v^*)^{\mathrm{T}}\xi, \quad 0 = A^{\mathrm{T}}\xi \\
&0 = \nabla_v \bar{O}(w^*, v^*) + \beta - \mu^1 - \mu^2 + \nabla_v \xi^{\mathrm{T}}(u^* - \Psi(v^*)w^*) \\
&0 \in \partial_u(-\beta^{\mathrm{T}}\sigma(u^*)) + \mu^1 + \alpha\mu^2 + \xi \\
&(\mu^1)^{\mathrm{T}}(v^* - u^*) = 0, \ \ (\mu^2)^{\mathrm{T}}(v^* - \alpha u^*) = 0
\end{aligned}
$$

Lemma 2

MFCQ [a] holds for the constraints set of problem (PP).

[a] [Mangasarian 1994]

- $\Rightarrow$ (w^*, b^*, v^*, u^*) is a l-stationary point of (PP) if and only if (w^*, b^*, v^*, u^*) is a KKT point of (PP)
- $\Rightarrow$ any local minimizer of (PP) is its l-stationary point

Main Theoretical Results

Let $\theta > \frac{1}{N}\|X\|_F^2$ and

$$\Omega_\theta = \{(w,b,v,u) : v-u \geq 0, v-\alpha u \geq 0,\ u = \Psi(v)w + Ab, \mathcal{O}(w,v,u) \leq \theta\}$$

Theorem 3

The set Ω_θ is bounded and the solution set to problem (PP) is nonempty and bounded.

(PP) :	global(local) minimizers	I-stationary point $\Leftrightarrow$	KKT point
	$\Downarrow\Downarrow$	some conditions $\Downarrow$	$\Downarrow$
(P) :	global(local) minimizers	MPCC C-stationary point	MPCC W-stationary point

- $\beta_\ell > LL_{\bar{\mathcal{O}}} \max\{\theta_w, 1\}^L + 2\sum_{j=\ell+1}^{L} \beta_j \theta_w \max\{\theta_w, 1\}^{j-\ell-1}$, $L_{\bar{\mathcal{O}}}$ is the Lipschitz modulus of $\bar{\mathcal{O}}$ over Ω_θ
- $\mathcal{O}(w,v,u) < \theta$

Main Theoretical Results

Let $\theta > \frac{1}{N}\|X\|_F^2$ and

$$\Omega_\theta = \{(w,b,v,u) : v-u \geq 0, v-\alpha u \geq 0,\ u = \Psi(v)w + Ab, \mathcal{O}(w,v,u) \leq \theta\}$$

Theorem 3

The set Ω_θ is bounded and the solution set to problem (PP) is nonempty and bounded.

(PP) :	global(local) minimizers	l-stationary point $\Leftrightarrow$	KKT point
	$\Downarrow\Downarrow$	some conditions $\Downarrow$	$\Downarrow$
(P) :	global(local) minimizers	MPCC C-stationary point	MPCC W-stationary point

- $\beta_\ell > LL_{\bar{\mathcal{O}}} \max\{\theta_w, 1\}^L + 2\sum_{j=\ell+1}^{L} \beta_j \theta_w \max\{\theta_w, 1\}^{j-\ell-1}$, $L_{\bar{\mathcal{O}}}$ is the Lipschitz modulus of $\bar{\mathcal{O}}$ over Ω_θ
- $\mathcal{O}(w,v,u) < \theta$

Extensions to ReLU Networks

$\alpha = 0 \Rightarrow$ the solution set of (PP) maybe unbounded

$\Downarrow$

solve (PP) over a constructed set

$$\begin{aligned} &\min_{w,b,v,u} \ \mathcal{O}(w,v,u) \\ &\text{s. t.} \ \ v-u \geq 0, v-\alpha u \geq 0, \ u = \Psi(v)w + Ab \\ &\qquad\ \ b \geq -e_{\overline{N}}\overline{N}(\theta_w + \theta_v) \end{aligned} \tag{PP$_\mathrm{b}$}$$

- the solution set to $(\mathrm{PP_b})$ is nonempty and bounded
- any local minimizer of (PP$_\mathrm{b}$) is its l-stationary point, and a MPCC-W stationary point of problem (P)
- the designed algorithm can be used to solve ReLU networks

3. Algorithm Framework and Convergence Analysis

An Inexact Augmented Lagrangian Method

[Lu-Zhang 2012; Chen et al. 2017]

Augmented Lagrangian function

$$\mathcal{L}_\rho(w,b,v,u;\xi) := \mathcal{O}(w,v,u) + \langle \xi, u - \Psi(v)w - Ab \rangle + \frac{\rho}{2}\|u - \Psi(v)w - Ab\|^2$$

- $\rho \in \mathbb{R}_+$
- $\xi \in \mathbb{R}^m$

Subproblem

$$\min_{(w,b,v,u): v \geq u,\, v \geq \alpha u} \mathcal{L}_\rho(w,b,v,u;\xi) \tag{4}$$

- IALAM： IALM framework with subproblem solved by an Alternating Minimization method

Algorithm Framework

The inexact augmented Lagrangian method for solving problem (PP) (IALM framework)

1. Input: $(w^{(0)}, b^{(0)}, v^{(0)}, u^{(0)}) \in \Omega_\theta$, $\rho^{(0)} > 0$, $\eta_1 \in (0,1)$, $\eta_2, \eta_3 > 1$, $\xi^{(0)} \in \mathbb{R}^m$, $\gamma \in \mathbb{N}_+$. $k := 1$
2. let $(\xi, \rho) = (\xi^{(k-1)}, \rho^{(k-1)})$, solve (4) inexactly
3. Calculate $\xi^{(k)} := \xi^{(k-1)} + \rho^{(k-1)}(u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)})$
4. if $k \le \gamma$, let $\rho^{(k)} = \rho^{(k-1)}$. If $k > \gamma$ and
$$\left\|u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)}\right\| \le \eta_1 \max_{t=k-\gamma,\dots,k-1} \left\|u^{(t)} - \Psi(v^{(t)})w^{(t)} - Ab^{(t)}\right\|$$
let $\rho^{(k)} = \rho^{(k-1)}$. otherwise, let
$$\rho^{(k)} = \max\left\{\rho^{(k-1)}/\eta_2, \|\xi^{(k)}\|^{1+\eta_3}\right\}$$
5. Let $k := k+1$, if stop criterion is not met, return to step 2.
6. Output: $(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})$

- the feasibility violation decreases non-monotonic.

Alternating Minimization Method for Solving (4)

Main Idea: solve (4) by splitting it into (w, b) and (v, u) blocks

why?

- the variable dimension of (4) is large
- w, b are variables of (1), meanwhile v, u are the auxiliary variables
- restricted to both of these two blocks are easy to solve
    - the (w, b) subproblem is strongly convex
    - the (v, u) subproblem has a closed-form unique solution

Alternating Minimization Method for Solving (4) (Cont'd)

Step 1: $(w^{(J)}, b^{(J)}) \rightarrow (w^{(J+1)}, b^{(J+1)})$

$$(w^{(J+1)}, b^{(J+1)}) := \arg\min_{w,b} \mathcal{L}_\rho(w, b, v^{(J)}, u^{(J)}; \xi)^1 \tag{5}$$

Step 2: $(v^{(J)}, u^{(J)}) \rightarrow (v^{(J+1)}, u^{(J+1)})$

$$(v^{(J+1)}, u^{(J+1)}) := \arg \min_{(v,u): v \geq u, v \geq \alpha u} \mathcal{L}_\rho(w^{(J+1)}, b^{(J+1)}, v, u; \xi) + \mathcal{P}(u, v; u^{(J)}, v^{(J)}, \tau^{(J)}) \tag{6}$$

$$\mathcal{P}(u, v; u^{(J)}, v^{(J)}, \tau^{(J)}) := \frac{1}{2} \sum_{n=1}^{N} \sum_{\ell=2}^{L} \left\| \begin{pmatrix} v_{n,\ell-1} \\ u_{n,\ell} \end{pmatrix} - \begin{pmatrix} v^{(J)}_{n,\ell-1} \\ u^{(J)}_{n,\ell} \end{pmatrix} \right\|^2_{S^{(J)}_\ell} + \frac{\tau_1}{2} \sum_{n=1}^{N} \|u_{n,1} - u^{(J)}_{n,1}\|^2$$

- $\tau_1 > 0$, $\tau^{(J)} := (\tau^{(J)}_2, \ldots, \tau^{(J)}_L)^{\mathrm{T}} \in \mathbb{R}^{L-1}$
- $S^{(J)}_\ell := \tau^{(J)}_\ell I_{N_\ell + N_{\ell-1}} - \rho \begin{bmatrix} -W^{(J+1)}_\ell & I_{N_\ell} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -W^{(J+1)}_\ell & I_{N_\ell} \end{bmatrix} \succeq \tau_1 I_{N_\ell + N_{\ell-1}}$
- $\tau^{(J)}_\ell := \rho \left\| \begin{bmatrix} -W^{(J+1)}_\ell & I_{N_\ell} \end{bmatrix} \right\|^2 + \tau_1$

[1] proximal gradient method [Dai-Fletcher 2005]

Alternating Minimization Method for Solving (4) (Cont'd)

Alternating Minimization Algorithm for Solving (4)

1. Input: $A, \xi, \rho > 0, (w^{(0)}, b^{(0)}, v^{(0)}, u^{(0)})$. let $\tau_1 > 0,\ \jmath = 0$
2. Update $(w^{(\jmath+1)}, b^{(\jmath+1)})$ by solving problem (6)
3. Update $(u^{(\jmath+1)}, v^{(\jmath+1)})$ by solving problem (6)
4. Set $\jmath := \jmath + 1$. If the stop criterion is not met, return to Step 2
5. Output: $(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)})$

Theorem 4

Let $\{(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)})\}$ be the sequence generated by the Alternating Minimization Algorithm. Then any accumulation point (w^*, b^*, v^*, u^*) of $\{(w^{(\jmath)}, b^{(\jmath)}, v^{(\jmath)}, u^{(\jmath)})\}$ is a KKT point of (4).

Convergence Analysis

Theorem 5

Let $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ *be the sequence generated by IALAM with* $\eta_3 > 1$*. Then the following statements hold.*

(a) $\liminf_{k\to\infty} \|u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)}\| = 0$ *and the sequence* $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ *has at least one accumulation point.*

(b) $\liminf_{k\to\infty} \mathrm{dist}((w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)}), \mathcal{Z}^*) = 0$*, where* $\mathcal{Z}^*$ *is the set of KKT points of (PP).*

(c) *If in addition that* $\gamma = 1$*, then* $\lim_{k\to\infty} \|u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)}\| = 0$*. Furthermore, any accumulation point* (w^*, b^*, v^*, u^*) *of* $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ *is a KKT point of problem (PP).*

- Theoretical Contribution: Different from the existing methods [Lu-Zhang 2012; Chen et al. 2017],we prove the existence of the accumulation point. Moreover, our designed algorithm support $\gamma > 1$
- Extensions: IALAM can be applied to many kind of networks

Convergence Analysis

Theorem 5

Let $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ *be the sequence generated by IALAM with* $\eta_3 > 1$. *Then the following statements hold.*

(a) $\liminf_{k\to\infty} \|u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)}\| = 0$ *and the sequence* $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ *has at least one accumulation point.*

(b) $\liminf_{k\to\infty} \mathrm{dist}((w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)}), \mathcal{Z}^*) = 0$, *where* $\mathcal{Z}^*$ *is the set of KKT points of (PP).*

(c) *If in addition that* $\gamma = 1$, *then* $\lim_{k\to\infty} \|u^{(k)} - \Psi(v^{(k)})w^{(k)} - Ab^{(k)}\| = 0$. *Furthermore, any accumulation point* (w^*, b^*, v^*, u^*) *of* $\{(w^{(k)}, b^{(k)}, v^{(k)}, u^{(k)})\}$ *is a KKT point of problem (PP).*

- Theoretical Contribution: Different from the existing methods [Lu-Zhang 2012; Chen et al. 2017],we prove the existence of the accumulation point. Moreover, our designed algorithm support $\gamma > 1$
- Extensions: IALAM can be applied to many kind of networks

4. Numerical Experiments

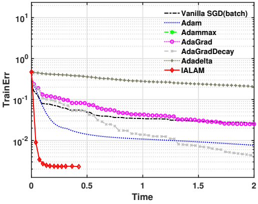
Default settings

- Stop criterion: $\rho^{(k)} > 10^3\rho^{(0)}$
- Initialization: $W_\ell^{(0)} = \mathrm{randn}(N_\ell, N_\ell - 1)/N$, $b^{(0)} = 0$, $u_{n,\ell}^{(0)} = W_\ell^{(0)} v_{n,\ell-1}^{(0)}$, $v_{n,\ell}^{(0)} = \sigma(u_{n,\ell}^{(0)})$
- Test problem: randomly generated synthetic dataset and MNIST dataset
- $N_{\text{test}} = \lceil N/5 \rceil$

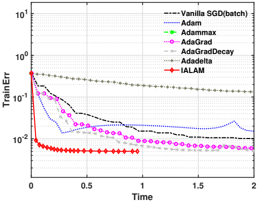
Output evaluation.

- TrainErr $= \frac{1}{N}\sum_{n=1}^{N} \|\sigma(W_L\sigma(\cdots\sigma(W_1x_n + b_1) + b_2\cdots) + b_L) - y_n\|^2$
- TestErr $= \frac{1}{N}\sum_{n=N+1}^{N+N_{\text{test}}} \|\sigma(W_L\sigma(\cdots\sigma(W_1x_n + b_1) + b_2\cdots) + b_L) - y_n\|^2$
- FeasVi $=$ $\frac{1}{N}\sum_{n=1}^{N}\sum_{\ell=1}^{L} \|v_{n,\ell} - \sigma(u_{n,\ell})\|^2 + \frac{1}{N}\sum_{n=1}^{N}\sum_{\ell=1}^{L} \|u_{n,\ell} - (W_\ell v_{n,\ell-1} + b_\ell)\|^2$
- Column Sparse Ratio: the ratio of columns in all the matrices W_ℓ, totaling $\sum_{\ell=0}^{L-1} N_\ell$ columns, where the l_2 norm values are below the tolerance ϵ.
- Accuracy (for the training set) and TestAcc

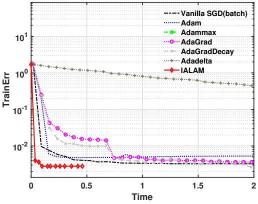
Comparisons Among IALAM and SGD-based Approaches on the Synthetic Dataset



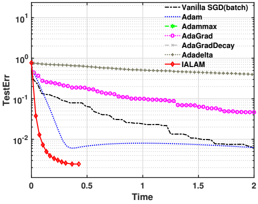
(a) $L = 2, N_1 = 10$



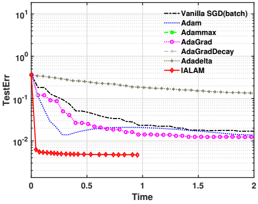
(b) $L = 3, N_1 = N_2 = 5$



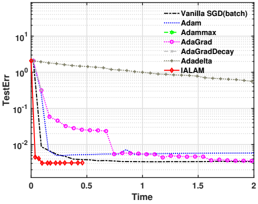
(c) $L = 4, N_1 = 4, N_2 = N_3 = 3$



(d) $L = 2, N_1 = 10$

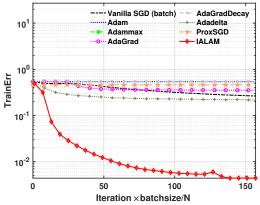


(e) $L = 3, N_1 = N_2 = 5$

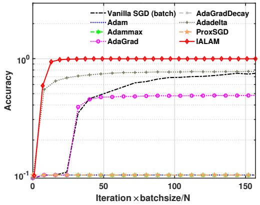


(f) $L = 4, N_1 = 4, N_2 = N_3 = 3$

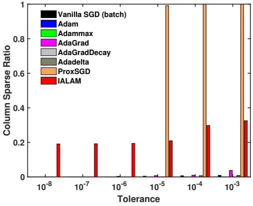
Comparisons Among IALAM and SGD-based Approaches on MNIST



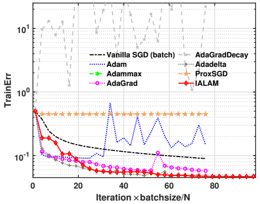
(a) TrainErr



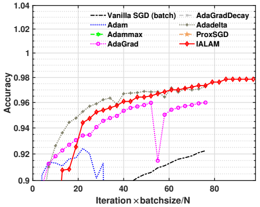
(b) Accuracy



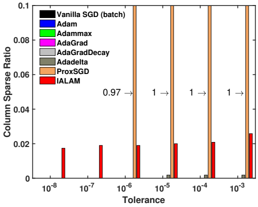
(c) Column Sparse Ratio



(d) TrainErr



(e) Accuracy



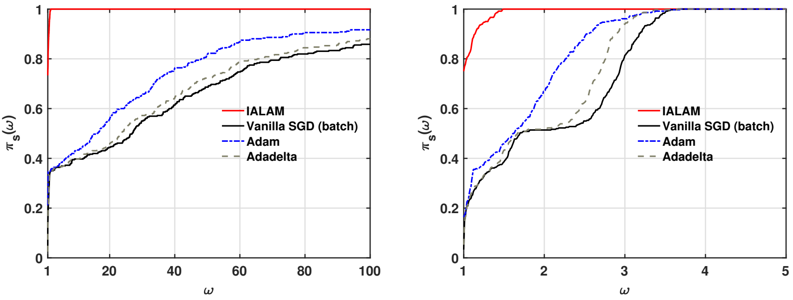
(f) Column Sparse Ratio

(a)–(c): $N = 1000, N_1 = 100, N_2 = 50, L = 3$

(d)–(f): $N = 60000, N_1 = 200, N_2 = 100, L = 3$

Overall Performance Profile on MNIST

[Dolan and More, 2002]



(a) TrainErr (b) TestErr

FIG 1: Performance profile for IALAM, Valinna SGD, Adadelta and Adam on TrainErr and TestErr.

We select 720 test problems based on MNIST data set with different network parameter combinations

Conclusion and Future Work

For the optimization problem (P) toward a nonconvex nonsmooth neuron network

- design a new model (PP) with bounded solution set
- present exact penalization：
  - local minmizers
  - l-stationary points
- design a IALAM algorithm for nonconvex problems.
  - converges to a KKT point/ l-stationary point of (PP)
  - the feasible violation decreases non-monotonic
  - the existence of approximation point is not required

Thanks for your listening!

Email: liuwei175@lsec.cc.ac.cn