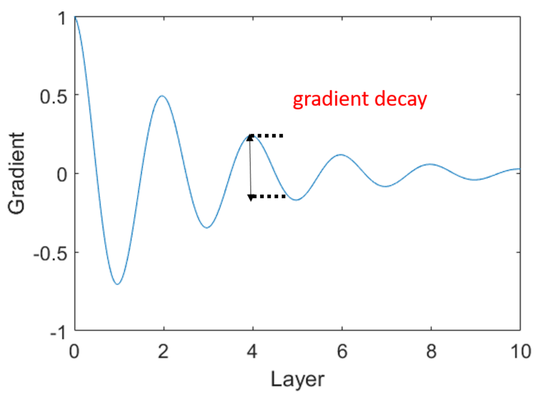
ADMM for Efficient Deep Learning with Global Convergence

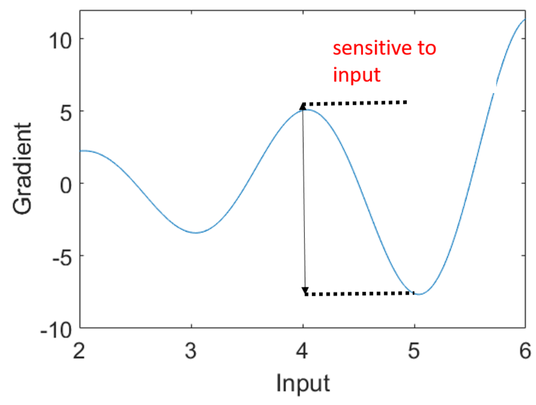
Junxiang Wang, Fuxun Yu, Xiang Chen and Liang Zhao

Cons of SGD as a Deep Learning Optimizer

SGD suffers from several limitations including



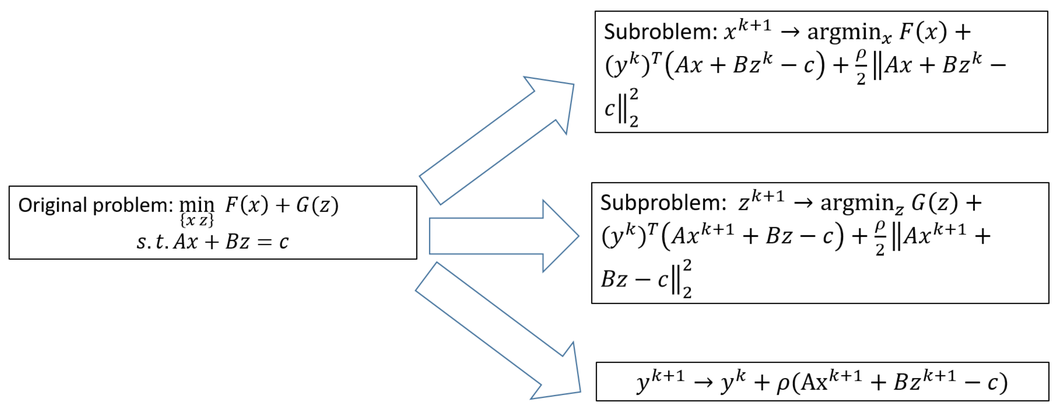
Gradient Vanishing



Poor Conditioning

Pros of ADMM as an Alternative

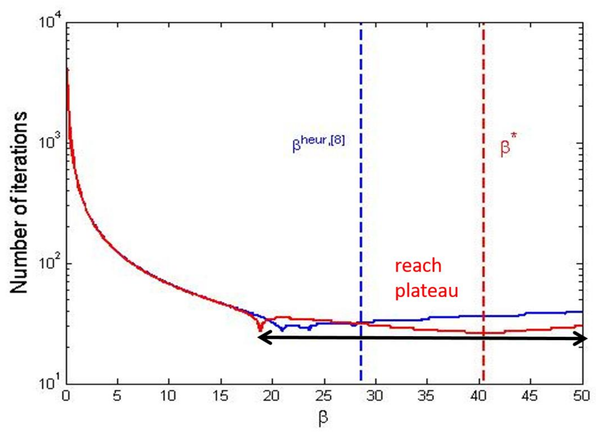
- The ADMM splits an optimization problem into two or more subproblems, each of which is easier to solve.



- The ADMM demonstrates excellent scalability in many deep learning applications.

Research Challenges

1. Slow convergence towards solutions.



2. Solving ADMM subproblems require computationally expensive matrix inversion.

$$A=\begin{bmatrix}2&1&1\\3&2&1\\2&1&2\end{bmatrix} \quad A^{-1}=?$$

3. The lack of global convergence guarantees.

Deep Learning ADMM (dlADMM)

Problem formulation is shown as follows:

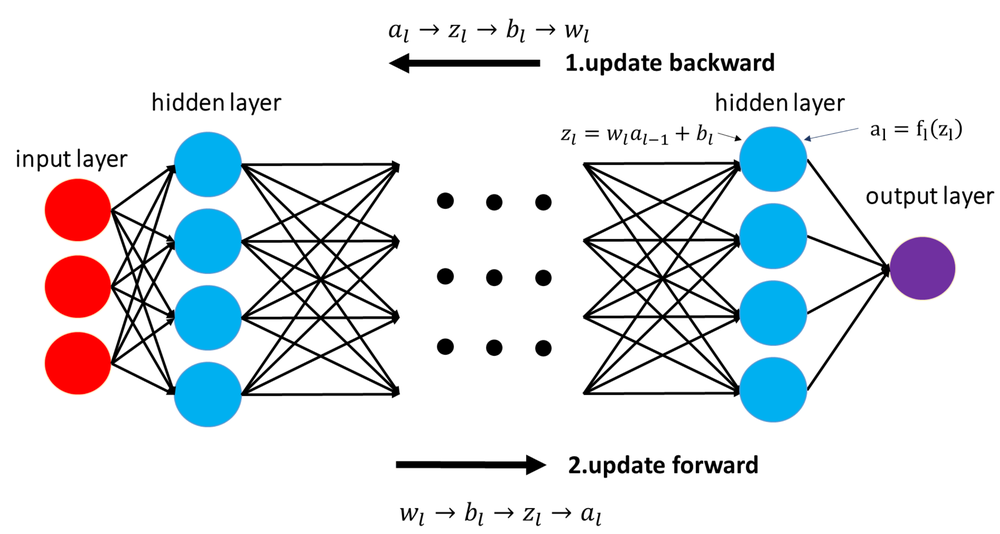
Problem 1.

$$\min_{W_l,b_l,z_l,a_l} \overbrace{R(z_L;y)}^{\text{loss function}}+\sum_{l=1}^{n}\overbrace{\Omega_l(W_l)}^{\text{regularization term}}$$
$$s.t.\ z_l=W_l a_{l-1}+b_l,(l=1,\cdots,L)$$
$$a_l=\underbrace{f_l(z_l)}_{\text{activation function}}(l=1,\cdots,L-1)$$

Problem 2.

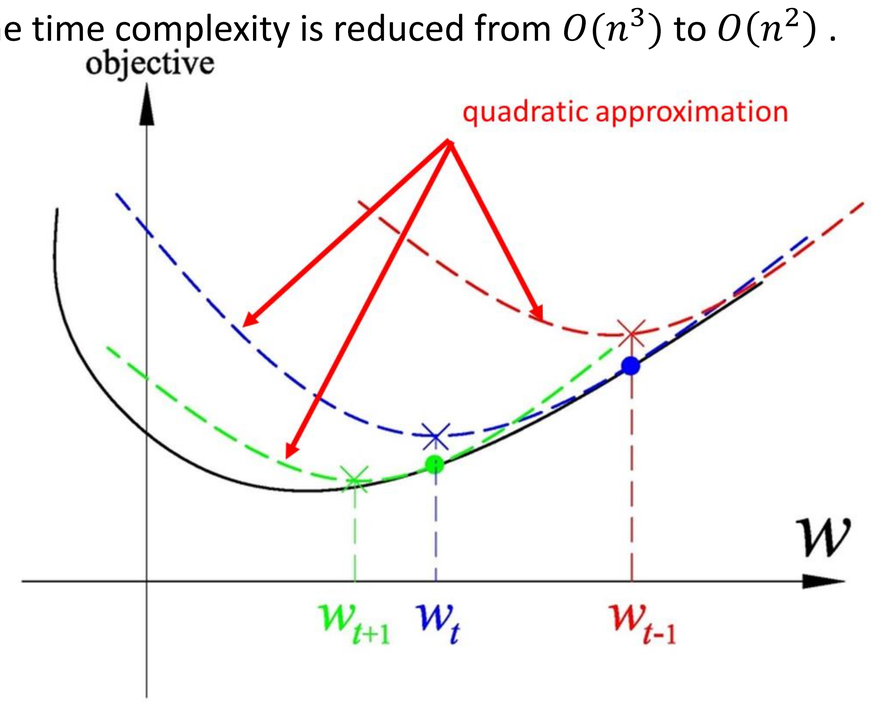
$$\min_{W_l,b_l,z_l,a_l} \overbrace{R(z_L;y)}^{\text{loss function}}+\sum_{l=1}^{n}\overbrace{\Omega_l(W_l)}^{\text{regularization term}}$$
$$\frac{\nu}{2}\sum_{l=1}^{L-1}(\|z_l-W_l a_{l-1}-b_l\|_2^2+\|a_l-\underbrace{f_l(z_l)}_{\text{activation function}}\|_2^2)$$
$$s.t.\ z_L=W_L a_{L-1}+b_L \text{ (the final layer)}$$

To accelerate convergence, the parameters are updated backward and then forward to exchange information efficiently.



Quadratic Approximation and Convergence Guarantees

To avoid matrix inversion, we apply the quadratic approximation techniques. In this way, the time complexity is reduced from $O(n^3)$ to $O(n^2)$.



The ADMM-type method (dlADMM) is firstly proven to converge to a critical point of deep learning problems, which means three properties:

Property 1: All variables are bounded and the objective is lower bounded.

Property 2: The objective decreases monotonically.

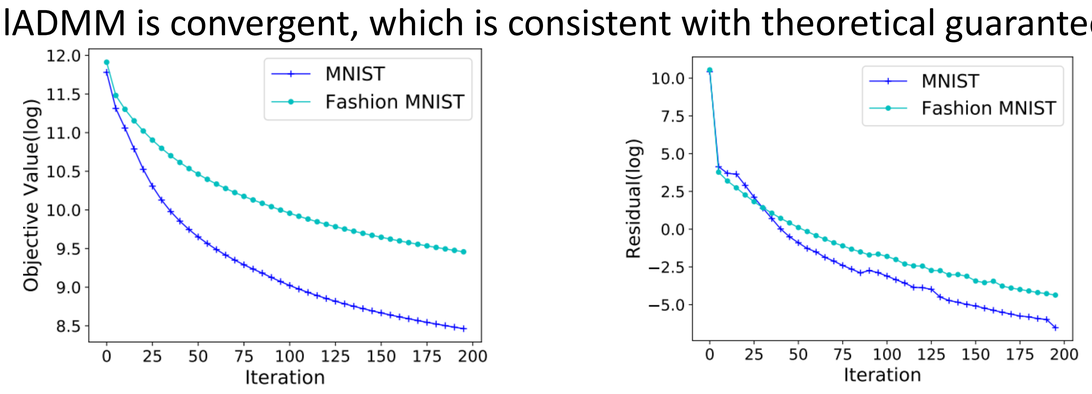
Property 3: The subgradient of the objective converges to 0.

The convergence rate of dlADMM is o$(\frac{1}{k})$.

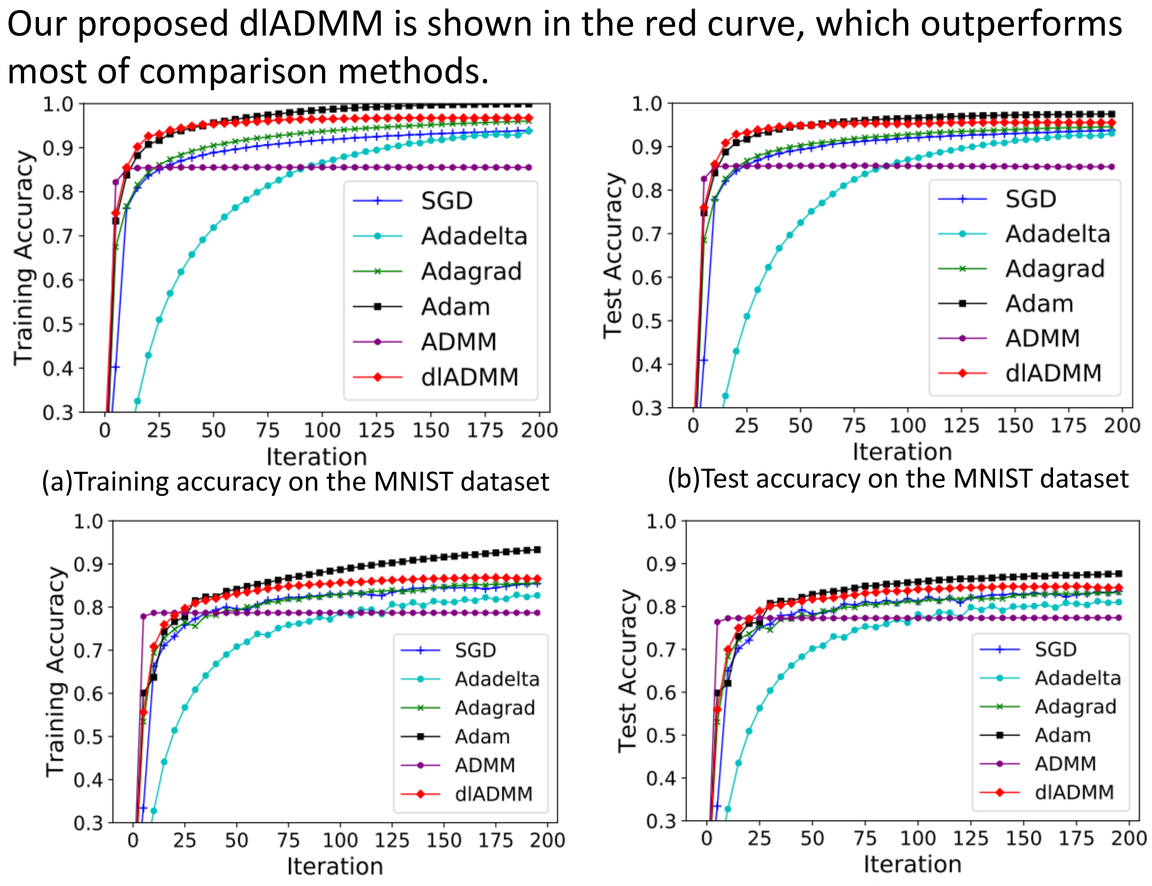
The code of our paper is available at
https://github.com/xianggebenben/dlADMM

Experiments Results

Our dlADMM is convergent, which is consistent with theoretical guarantees.



Our proposed dlADMM is shown in the red curve, which outperforms most of comparison methods.



(a)Training accuracy on the MNIST dataset

(b)Test accuracy on the MNIST dataset

(c) Training accuracy on the Fashion-MNIST dataset

(d)Test accuracy on the Fashion-MNIST dataset