KDD2019

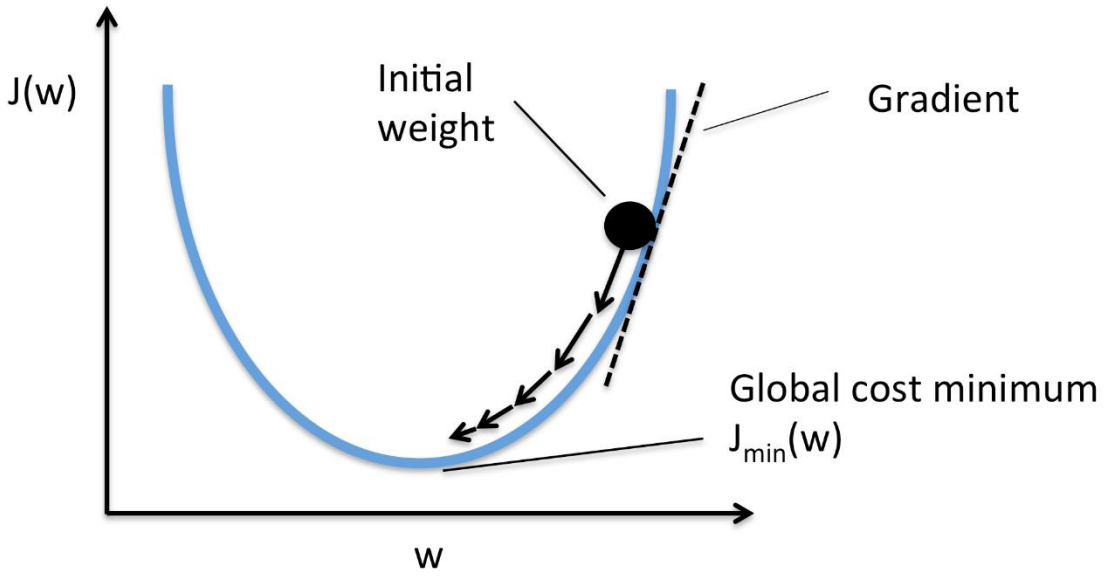
ADMM for Efficient Deep Learning with Global Convergence

Junxiang Wang, Fuxun Yu, Xiang Chen and Liang Zhao

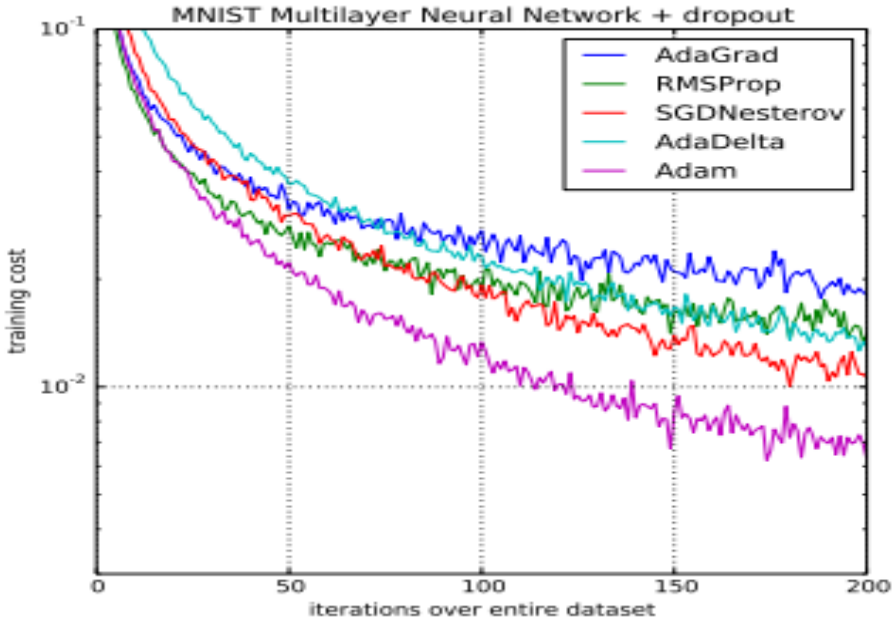
Presented by Junxiang Wang

SGD and Its Variants as Deep Learning Optimizers

Stochastic gradient descent(SGD) and its variants are state-of-the-art optimization methods in deep learning problems.



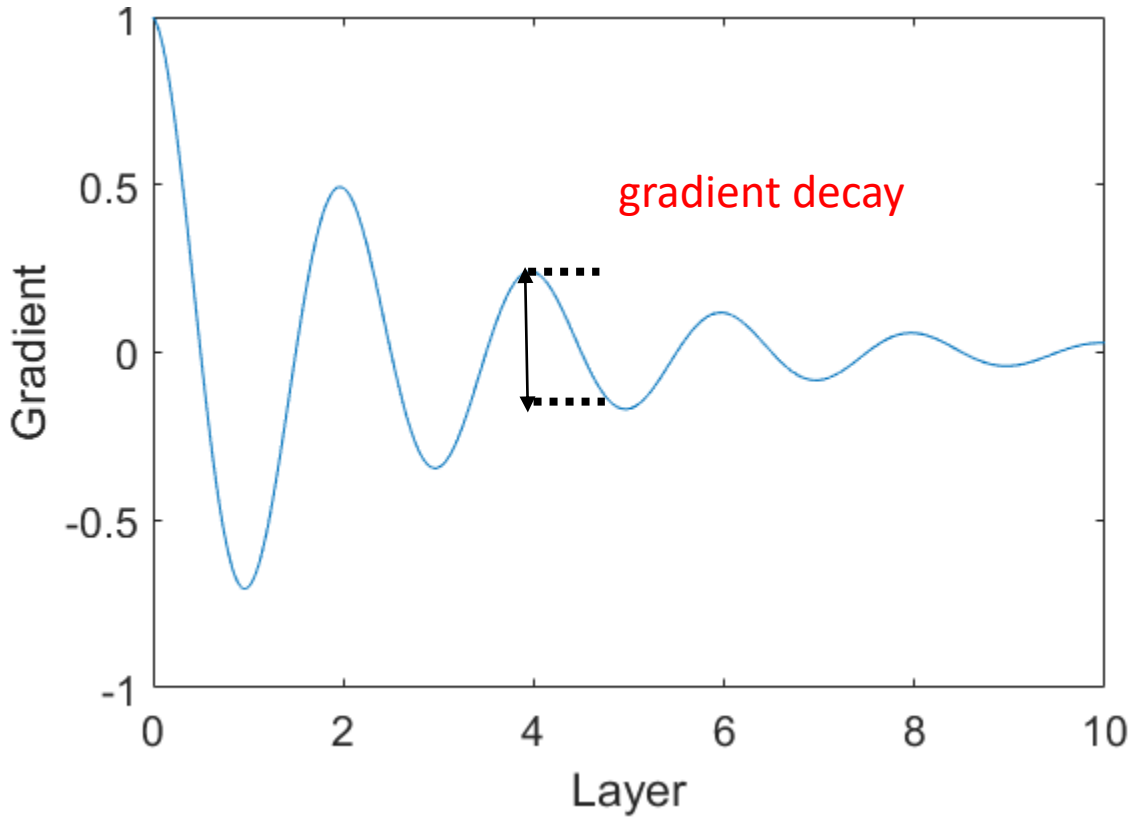
Stochastic gradient descent(SGD)



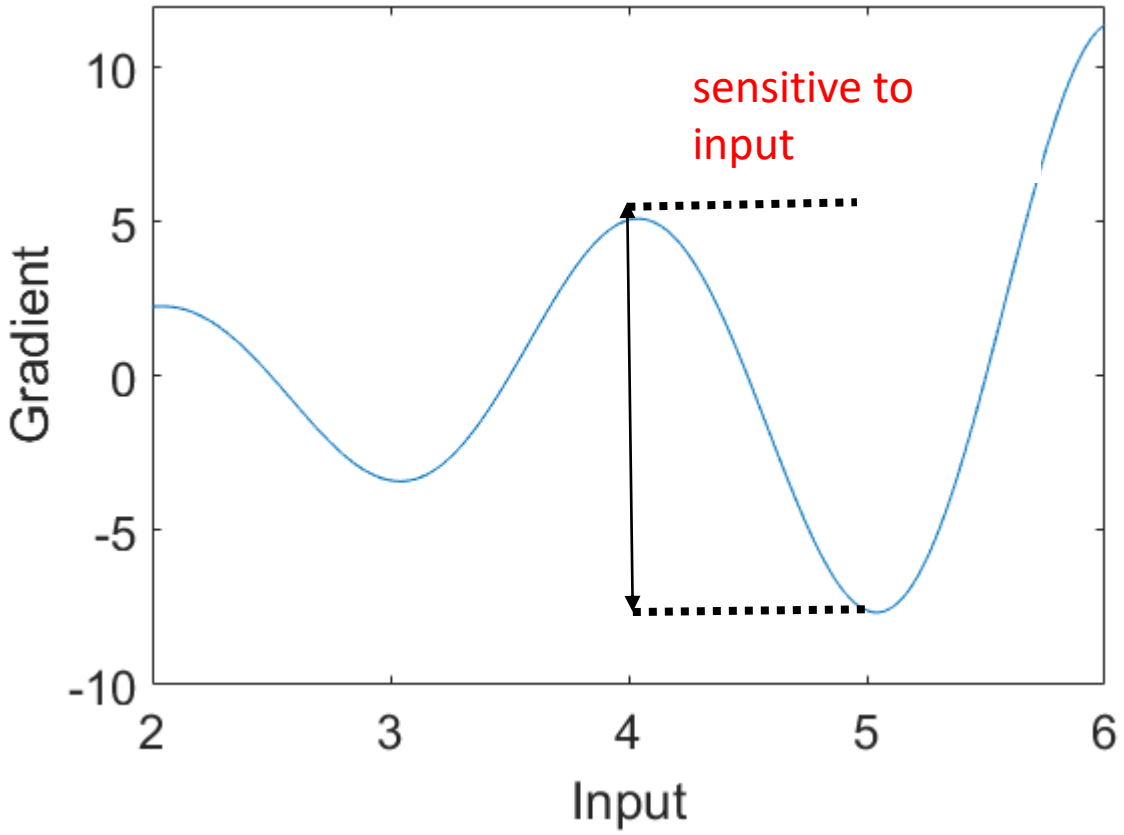
Outstanding Performance

SGD and Its Variants as Deep Learning Optimizers

However, SGD suffers from several limitations including



Gradient Vanishing



Poor Conditioning

ADMM as an Alternative

Recently, the Alternating Direction Method of Multipliers (ADMM) has been widely used in many deep learning applications.

Original problem: $\min_{\{x\ z\}} F(x) + G(z)$
$s.t. Ax + Bz = c$

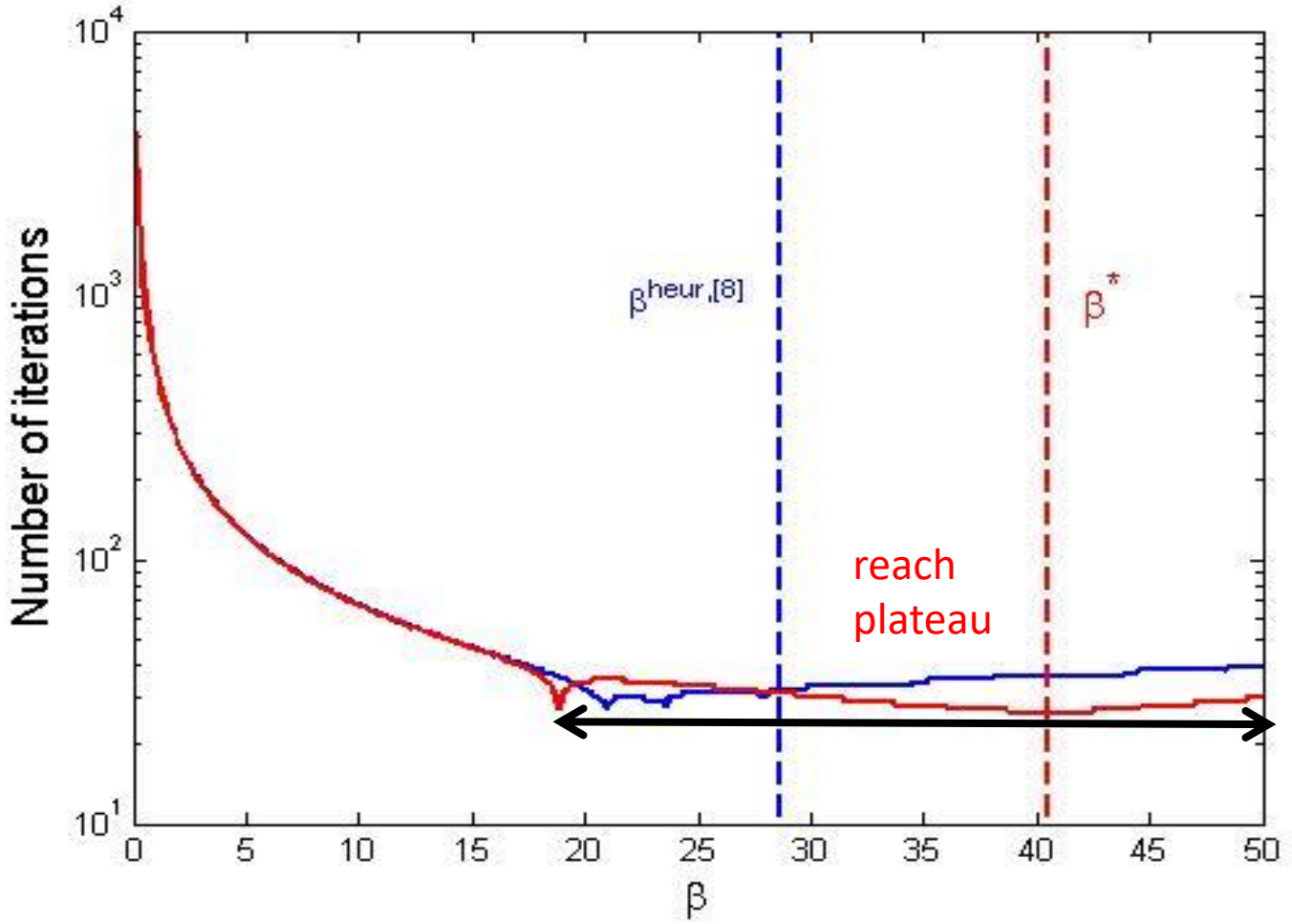
Subroblem: $x^{k+1} \rightarrow \text{argmin}_x F(x) + (y^k)^T\left(Ax + Bz^k - c\right) + \frac{\rho}{2}\left\|Ax + Bz^k -\right.$

Subproblem: $z^{k+1} \rightarrow \text{argmin}_z G(z) + (y^k)^T\left(Ax^{k+1} + Bz - c\right) + \frac{\rho}{2}\left\|Ax^{k+1} +\right.$

$y^{k+1} \rightarrow y^k + \rho(\text{Ax}^{k+1} + Bz^{k+1} - c)$

The Challenges of ADMM for Deep Learning Problems

1. Slow convergence towards solutions.



The Challenges of ADMM for Deep Learning Problems

2. Solving ADMM subproblems require matrix inversion, which is $O(n^3)$.

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

3. The lack of global convergence guarantees.

KDD2019

Deep Learning Alternating Direction Method of Multipliers(dlADMM)

Problem 1.

$$\min_{W_l,b_l,z_l,a_l} \overbrace{R(z_L;y)}^{\text{loss function}} + \overbrace{\sum_{l=1}^{n}\Omega_l(W_l)}^{\text{regularization term}}$$

$$s.t.\, z_l = W_l a_{l-1} + b_l, (l = 1, \cdots, L)$$

$$a_l = \underbrace{f_l(z_l)}_{\text{activation function}} (l = 1, \cdots, L-1)$$

Problem 2.

$$\min_{W_l,b_l,z_l,a_l} \overbrace{R(z_L;y)}^{\text{loss function}} + \overbrace{\sum_{l=1}^{n}\Omega_l(W_l)}^{\text{regularization term}}$$

$$\frac{\nu}{2}\sum_{l=1}^{L-1}(\|z_l - W_l a_{l-1} - b_l\|_2^2 + \|a_l - \underbrace{f_l(z_l)}_{\text{activation function}}\|_2^2)$$

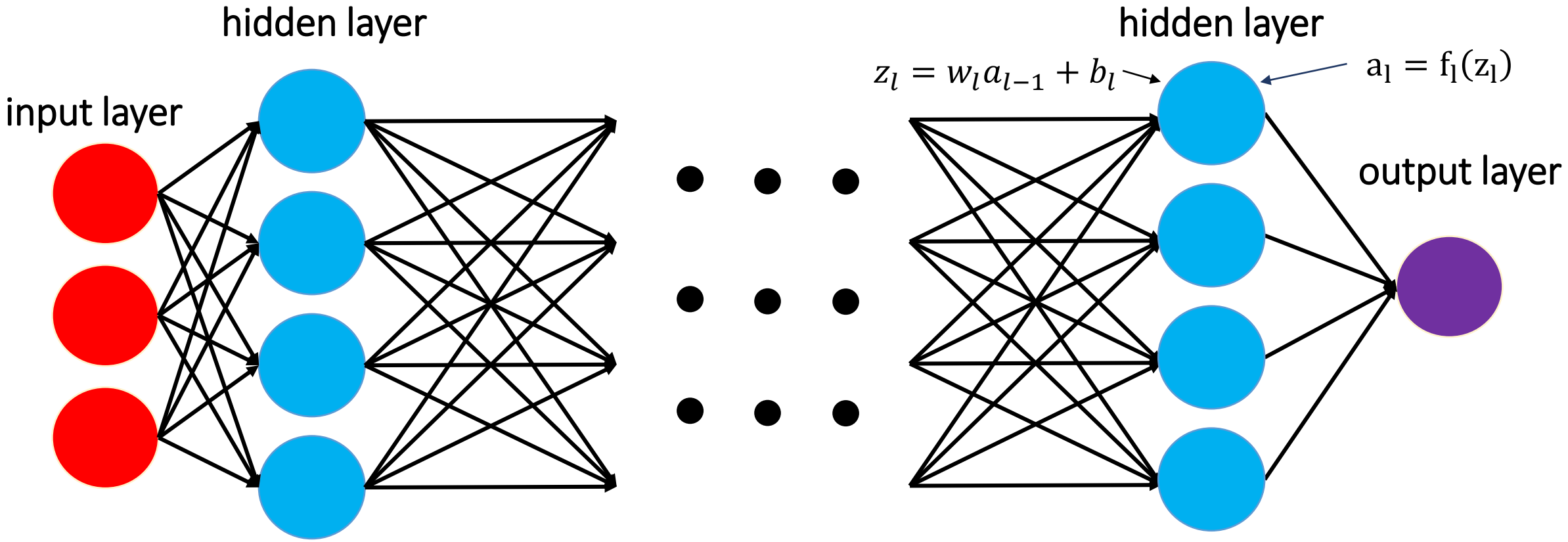
$$s.t.\, z_L = W_L a_{L-1} + b_L \text{ (the final layer)}$$

KDD2019

dlADMM

$a_l \rightarrow z_l \rightarrow b_l \rightarrow w_l$

1.update backward

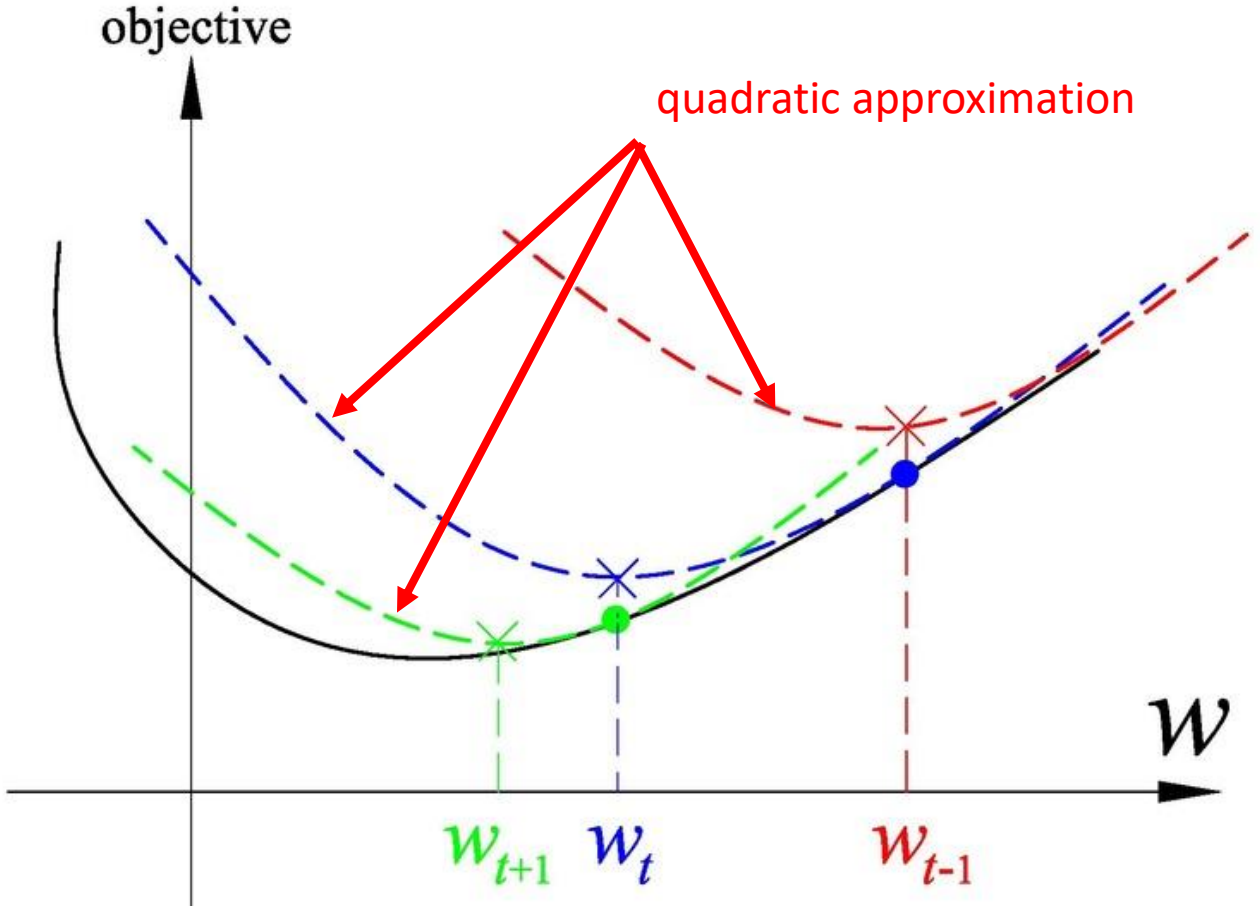


2.update forward

$w_l \rightarrow b_l \rightarrow z_l \rightarrow a_l$

dlADMM

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



dlADMM

We give the first proof that our dlADMM converges to a critical point of Problem 2. Which means that:

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

Property 2: The objective decreases monotonically.

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

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

See the paper for details.

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