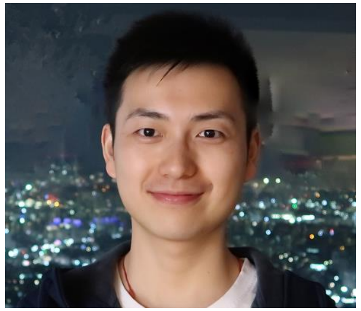
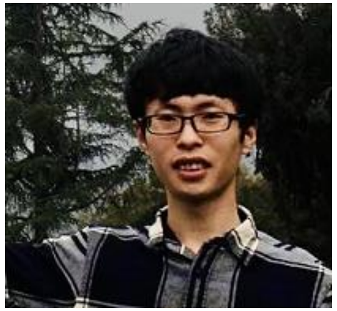
ICML | 2019

Thirty-sixth International
Conference on Machine Learning

Differentiable Linearized ADMM





Xingyu Xie[*, 1] Jianlong Wu[*, 1] Zhisheng Zhong[1] <u>Guangcan Liu</u>[✉, 2] Zhouchen Lin[✉, 1]

[1] Key Lab. of Machine Perception, School of EECS, Peking University
[2] B-DAT and CICAEET, School of Automation, Nanjing University of Information Science and Technology

Background

- Optimization plays a very important role in learning
  - Most machine learning problems are, in the end, optimization problems
    - SVM
    - K-Means
    - …
    - Deep Learning

$$\boxed{\min_{x}\ f(x, \text{data}), \qquad s.t. \quad x \in \Theta}$$

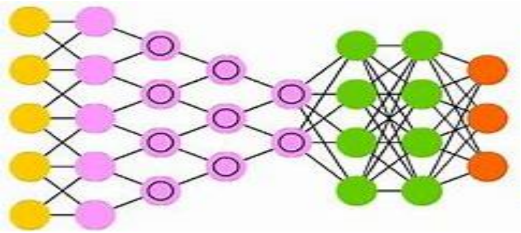
--- personal opinions: In general, what the computers can do is nothing more than “computation”. Thus, to assign them the ability to “learn”, it is often desirable to convert a “learning” problem into some kind of computational problem.

- Question: Conversely, can optimization benefit from learning ?

Learning-based Optimization

- A traditional optimization algorithm is indeed an ultra-deep network with fixed parameters

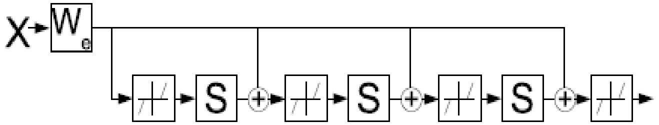
$$\min_{x} \ f(x, \text{data}), \qquad s.t. \quad x \in \Theta \qquad\qquad x_{t+1} = g(x_t)$$

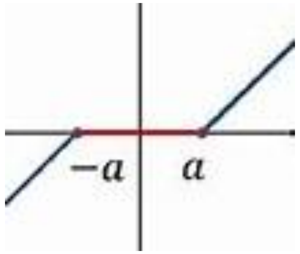


$$\min_{x} \parallel y - Ax \parallel_2^2 + \lambda \parallel x \parallel_1$$

$$x_{t+1} = h_\theta(W_e y + S x_t)$$

$$S = I - \frac{A^T A}{\rho}, W_e = \frac{A^T}{\rho}$$





- Learning-based optimization: Introduce learnable parameters and “reduce” the network depth, so as to improve computational efficiency

- Gregor K, Lecun Y. Learning fast approximations of sparse coding. ICML 2010.
- P. Sprechmann, A. M. Bronstein, and G. Sapiro Learning, *Efficient Sparse and Low Rank Models*, TPAMI 2015
- Yan Yang, Jian Sun, Huibin Li, Zongben Xu. ADMM-Net: A deep learning approach for compressive sensing MRI, NeurIPS 2016.
- Brandon Amos, J. Zico Kolter. OptNet: optimization method as a layer in neural network. ICML 2017.

Learning-based Optimization (Con’t)

- Limits of existing work
  - In a theoretical point of view, it is unclear why learning can improve computational efficiency, as theoretical convergence analysis is extremely rare
    - X. Chen, J. Liu, Z. Wang, W. Yin, Theoretical linear convergence of unfolded ISTA and its practical weights and thresholds, NeurIPS, 2018.

$$\operatorname*{minimize}_{x} \frac{1}{2}\|b - Ax\|_2^2 + \lambda\|x\|_1$$

    - specific to unconstrained problems

D-LADMM: Differentiable Linearized ADMM

Target constrained problem: $\min_{\mathbf{Z},\mathbf{E}} \; f(\mathbf{Z}) + g(\mathbf{E}), \quad \text{s.t. } \mathbf{X} = \mathbf{A}\mathbf{Z} + \mathbf{B}\mathbf{E},$

convex

known

LADMM (Lin et al, NeurIPS 2011):

$$\begin{cases} \mathbf{T}_{k+1} = \mathbf{A}\mathbf{Z}_k + \mathbf{B}\mathbf{E}_k - \mathbf{X}, \\ \mathbf{Z}_{k+1} = \mathbf{prox}_{\frac{f}{L_1}} \left\{ \mathbf{Z}_k - \frac{1}{L_1}\mathbf{A}^\top(\boldsymbol{\lambda}_k + \beta\mathbf{T}_{k+1}) \right\} \\ \widehat{\mathbf{T}}_{k+1} = \mathbf{A}\mathbf{Z}_{k+1} + \mathbf{B}\mathbf{E}_k - \mathbf{X}, \\ \mathbf{E}_{k+1} = \mathbf{prox}_{\frac{g}{L_2}} \left\{ \mathbf{E}_k - \frac{1}{L_2}\mathbf{B}^\top(\boldsymbol{\lambda}_k + \beta\widehat{\mathbf{T}}_{k+1}) \right\} \\ \boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \beta(\mathbf{A}\mathbf{Z}_{k+1} + \mathbf{B}\mathbf{E}_{k+1} - \mathbf{X}), \end{cases}$$

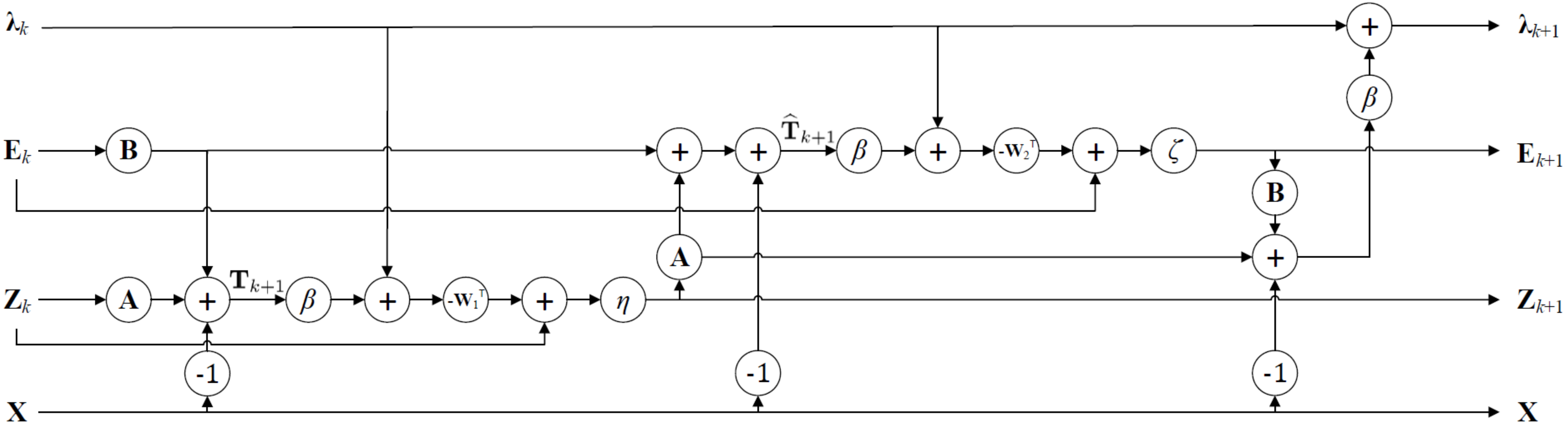
D-LADMM:

$$\begin{cases} \mathbf{T}_{k+1} = \mathbf{A}\mathbf{Z}_k + \mathbf{B}\mathbf{E}_k - \mathbf{X}, \\ \mathbf{Z}_{k+1} = \eta_{(\boldsymbol{\theta}_1)_k}\left(\mathbf{Z}_k - (\mathbf{W}_1)_k^\top(\boldsymbol{\lambda}_k + \boldsymbol{\beta}_k \circ \mathbf{T}_{k+1})\right), \\ \widehat{\mathbf{T}}_{k+1} = \mathbf{A}\mathbf{Z}_{k+1} + \mathbf{B}\mathbf{E}_k - \mathbf{X}, \\ \mathbf{E}_{k+1} = \zeta_{(\boldsymbol{\theta}_2)_k}\left(\mathbf{E}_k - (\mathbf{W}_2)_k^\top(\boldsymbol{\lambda}_k + \boldsymbol{\beta}_k \circ \widehat{\mathbf{T}}_{k+1})\right), \\ \boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \boldsymbol{\beta}_k \circ (\mathbf{A}\mathbf{Z}_{k+1} + \mathbf{B}\mathbf{E}_{k+1} - \mathbf{X}), \end{cases}$$

$\eta(\cdot)$ and $\zeta(\cdot)$ are learnable non-linear functions

learnable param.: $\Theta = \{(\mathbf{W}_1)_k, (\mathbf{W}_2)_k, (\boldsymbol{\theta}_1)_k, (\boldsymbol{\theta}_2)_k, \boldsymbol{\beta}_k\}_{k=0}^{K}$

D-LADMM (Con’t)



Questions:

Q1: Can D-LADMM guarantee to solve correctly the optimization problem?

Q2: What are the benefits of D-LADMM?

Q3: How to train the model of D-LADMM?

Main Assumption

- assumption required by LADMM:
- assumption required by D-LADMM:

$$\begin{tikzcd}
\frac{1}{t}\mathbf{I} - \mathbf{A}^{\top}\mathbf{A} \succ 0 \arrow[r, "\text{generalized}"] & \text{none-emptiness of } \underbrace{\mathcal{S}(\sigma, \mathbf{A}) := \left\{ (\mathbf{W}, \boldsymbol{\theta}, \boldsymbol{\beta}) \middle| \|\mathbf{W} - \mathbf{A}\| \leq \sigma, \mathcal{D} \succ 0, \boldsymbol{\beta}, \boldsymbol{\theta} > 0 \right\}}_{\text{Assumption 1}} \arrow[l, bend left, "{\mathbf{W} = \mathbf{A},\ \boldsymbol{\theta} = \frac{1}{t} \quad \text{and} \quad \boldsymbol{\beta} = \mathbf{1}}"]
\end{tikzcd}$$

Theoretical Result I

Q1: Can D-LADMM guarantee to solve correctly the optimization problem?

A1: Yes!

$$\boxed{\boldsymbol{\omega}_k := (\mathbf{Z}_k, \mathbf{E}_k, -\boldsymbol{\lambda}_k)}$$

D-LADMM’s k-th layer output

$$\boxed{\boldsymbol{\Omega}^*}$$

solution set of original problem

$$\boxed{\mathrm{dist}(\boldsymbol{\omega}, \boldsymbol{\Omega}^*)}$$

distance to the solution set

Theorem 1 and Theorem 2 [**Convergence and Monotonicity**] (informal).

$$\underbrace{\mathrm{dist}(\boldsymbol{\omega}_{k+1}, \boldsymbol{\Omega}^*) \geq \mathrm{dist}(\boldsymbol{\omega}_{k+1}, \boldsymbol{\Omega}^*) \to 0,}_{}\ \text{ as } k \to \infty.$$

$$\Downarrow$$

$$\boldsymbol{\omega}_k \to \boldsymbol{\omega}^* \in \boldsymbol{\Omega}^*$$

Theoretical Result II

Q2: What are the benefits of D-LADMM? A2: Converge faster!

D-LADMM > LADMM

Theorem 3 [**Convergence Rate**] (informal).
If the original problem satisfies *Error Bound Condition (condition on* **A** and **B***)*, then

$$\mathrm{dist}(\boldsymbol{\omega}_{k+1}, \boldsymbol{\Omega}^*) < \gamma \ \mathrm{dist}(\boldsymbol{\omega}_k, \boldsymbol{\Omega}^*), \quad \text{where } 0 < \gamma < 1.$$

⇒ **linear convergence**

General case (no EBC):

Lemma 4.4 [**Faster Convergence**] (informal).
Define operators: $\boldsymbol{\omega}_{k+1} := \mathcal{T}_{\Theta_k}(\boldsymbol{\omega}_k)$ for D-LADMM; $\boldsymbol{\omega}_{k+1} := \mathcal{T}(\boldsymbol{\omega}_k)$ for LADMM.
For any ω,

$$\mathrm{dist}(\mathcal{T}_{\Theta}(\boldsymbol{\omega}), \boldsymbol{\Omega}^*) \leq \mathrm{dist}(\mathcal{T}(\boldsymbol{\omega}), \boldsymbol{\Omega}^*).$$

Training Approaches

Q3: How to train the model of D-LADMM?

- Unsupervised way: minimizing duality gap

$$\min_{\Theta} f(\mathbf{Z}_K) + g(\mathbf{E}_K) - d^*(\boldsymbol{\lambda}_K),$$

where $d^*(\boldsymbol{\lambda}_K) = \inf_{\mathbf{Z},\mathbf{E}} f(\mathbf{Z}) + g(\mathbf{E}) + \langle \boldsymbol{\lambda}_K, \mathbf{AZ} + \mathbf{BE} - \mathbf{X} \rangle$ is the dual function.

Global optimum is attained whenever the objective (duality gap) reaches zero!

- Supervised way: minimizing square loss

$$\min_{\Theta} \|\mathbf{Z}_K - \mathbf{Z}^*\|_F^2 + \|\mathbf{E}_K - \mathbf{E}^*\|_F^2.$$

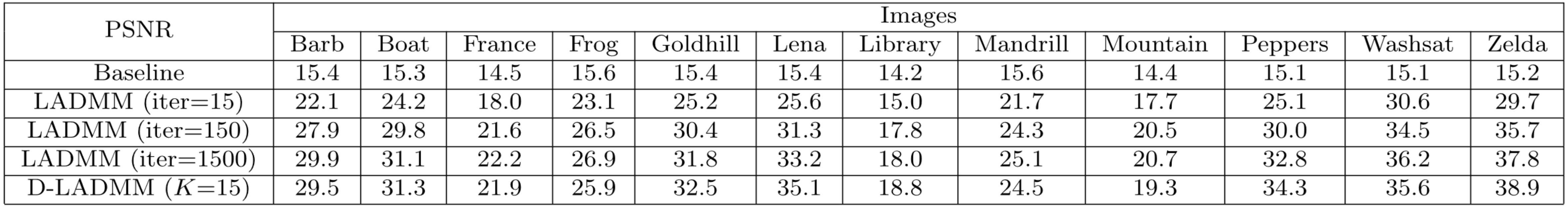
ground-truth Z^* and E^* are provided along with the training samples

Experiments

Target optimization problem

$$\min_{\mathbf{Z},\mathbf{E}} \ \lambda\|\mathbf{Z}\|_1 + \|\mathbf{E}\|_1, \qquad s.t. \ \mathbf{X} = \mathbf{AZ} + \mathbf{E}.$$

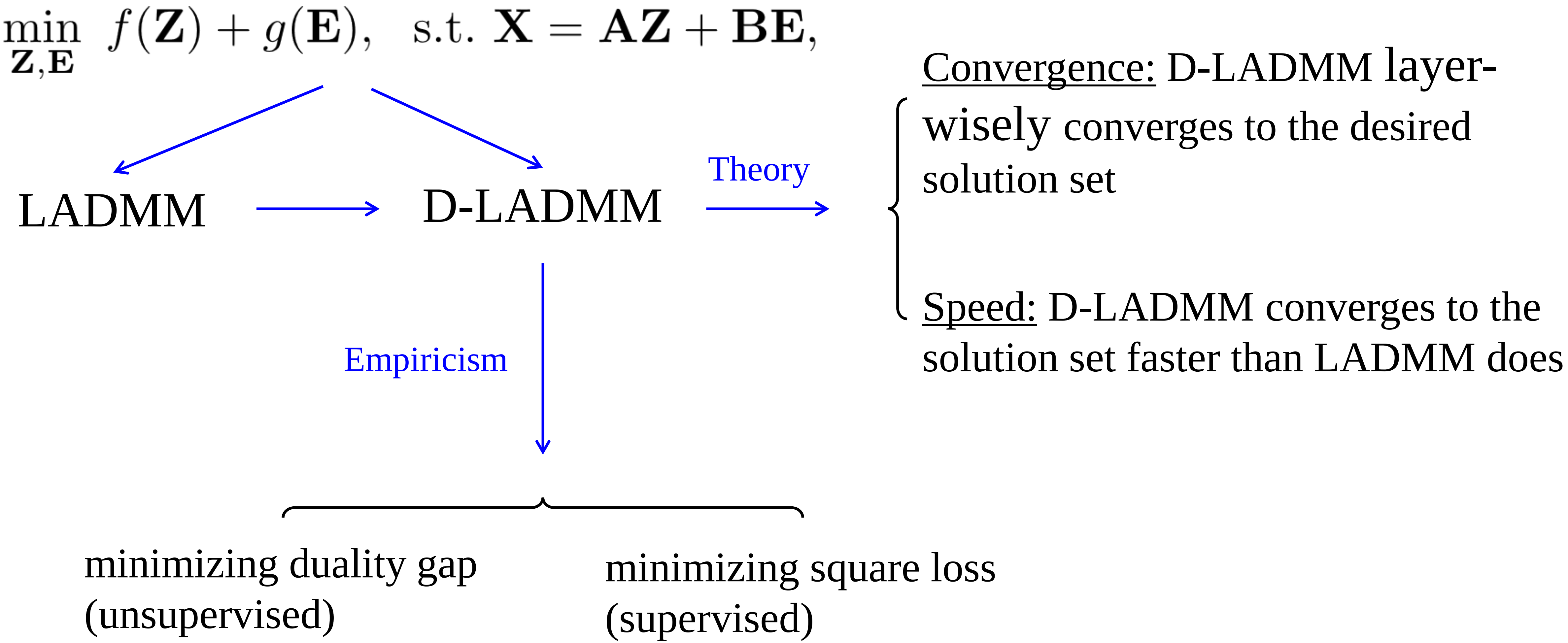
Table 1. PSNR comparison on 12 images with noise rate 10%.



PSNR	Images											
	Barb	Boat	France	Frog	Goldhill	Lena	Library	Mandrill	Mountain	Peppers	Washsat	Zelda
Baseline	15.4	15.3	14.5	15.6	15.4	15.4	14.2	15.6	14.4	15.1	15.1	15.2
LADMM (iter=15)	22.1	24.2	18.0	23.1	25.2	25.6	15.0	21.7	17.7	25.1	30.6	29.7
LADMM (iter=150)	27.9	29.8	21.6	26.5	30.4	31.3	17.8	24.3	20.5	30.0	34.5	35.7
LADMM (iter=1500)	29.9	31.1	22.2	26.9	31.8	33.2	18.0	25.1	20.7	32.8	36.2	37.8
D-LADMM (K=15)	29.5	31.3	21.9	25.9	32.5	35.1	18.8	24.5	19.3	34.3	35.6	38.9

15-layer D-LADMM achieves a performance comparable to, or even slightly better than, the LADMM algorithm with 1500 iterations!

Conclusion



$$\min_{\mathbf{Z},\mathbf{E}} \ f(\mathbf{Z}) + g(\mathbf{E}), \quad \text{s.t. } \mathbf{X} = \mathbf{AZ} + \mathbf{BE},$$

LADMM

D-LADMM

Theory

Convergence: D-LADMM layer-wisely converges to the desired solution set

Speed: D-LADMM converges to the solution set faster than LADMM does

Empiricism

minimizing duality gap (unsupervised)

minimizing square loss (supervised)