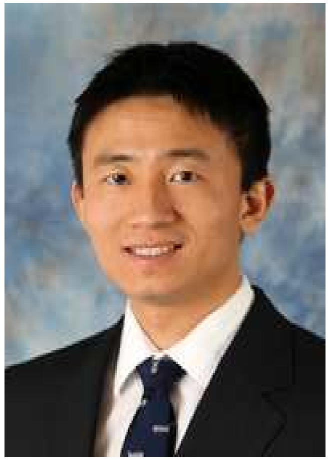
PA-GD: On the Convergence of Perturbed Alternating Gradient Descent to Second-Order Stationary Points for Structured Nonconvex Optimization

Presenter: Songtao Lu

University of Minnesota Twin Cities

Joint work with Mingyi Hong and Zhengdao Wang

Co-authors



Mingyi Hong

University of Minnesota

Zhengdao Wang

Iowa State University

Agenda

- **Motivation**
  - A class of structured non-convex problems
- **What we plan to achieve**:
  - **Random perturbation**:
    Convergence rate of alternating gradient descent (**A-GD**) to second-order stationary points (**SOSPs**) with high probability
- **Numerical Results**
  - Two-layer linear neural networks:
  - Matrix factorization
- **Concluding Remarks**

Block Structured Nonconvex Optimization

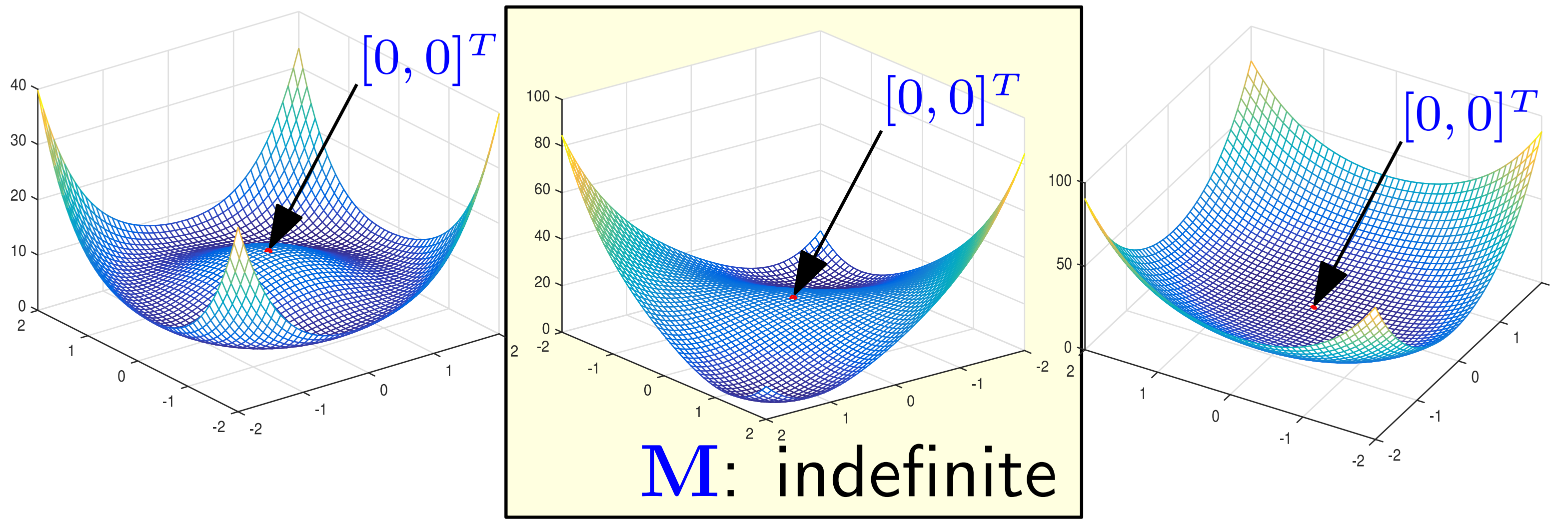
- Consider the following problem

$$\mathbb{P}: \qquad \underset{\mathbf{x},\mathbf{y}}{\text{minimize}} \quad f(\mathbf{x},\mathbf{y})$$

- $f(\mathbf{x},\mathbf{y})$: $\mathbb{R}^d \to \mathbb{R}$ is a smooth nonconvex function
  - $\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}$
  - $\mathbf{y} \in \mathbb{R}^{d_\mathbf{y}}$
  - $d = d_\mathbf{x} + d_\mathbf{y}$

Motivation: Nice Landscapes

- High dimensional problems: strict saddle points common
- There are some nice/benign block structured problems [R. Ge et al., 2017, J. Lee et al., 2018]
  - All local minima are global minima
  - Saddle points: very poor compared with local minima
  - Every saddle point: strict (Hessian matrix has at least one negative eigenvalue)



local maximum — strict saddle — local minimum

$$\underset{\mathbf{x}\in\mathbb{R}^{2\times 1}}{\text{minimize}} \|\mathbf{x}\mathbf{x}^T - \mathbf{M}\|_F^2$$

Optimality Conditions

- Common definition of first-order stationary points (FOSPs)

$$\|\nabla f(\mathbf{x}, \mathbf{y})\| \leq \epsilon$$

where $\epsilon > 0$, then $(\mathbf{x}, \mathbf{y})$ is an ϵ-FOSP.

- Common definition of SOSPs

If the following holds

$$\|\nabla f(\mathbf{x}, \mathbf{y})\| \leq \epsilon, \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(\mathbf{x}, \mathbf{y})) \geq -\gamma$$

where $\epsilon, \gamma > 0$, then $(\mathbf{x}, \mathbf{y})$ is an (ϵ, γ)-SOSP.

Literature

Algorithms with convergence guarantees to SOSPs:

- Second-order methods (one block)
  - Trust region method [Conn et al., 2000]
  - Cubic regularized Newton's method [Nesterov & Polyak, 2006]
  - Hybrid of first-order and second-order method [Reddi et al., 2018]
- First-order methods (one block)
  - Perturbed gradient descent (PGD) [Jin et al., 2017]
  - Stochastic first order method (NEgative-curvature-Originated-from-Noise, NEON, [Xu et al., 2017])
  - Neon2 (finding local minima via first-order oracles) [Allen-Zhu et al., 2017]
  - Accelerated methods [Carmon et al., 2016][Jin et al., 2018][Xu et al., 2018]
  - Many more

Literature

- Block structured nonconvex optimization (asymptotic) :
  - Block coordinate descent (BCD) [Song et al., 2017][Lee et al., 2017]
  - Alternating direction methods of multipliers (ADMM) [Hong et al., 2018]

- But none of these work has shown the convergence rate of block coordinate descent to SOSPs, even for the two-block case.

- Gradient descent can take exponential number of iterations to escape saddle points [Du et al., 2017]

Motivation: Block Structured Nonconvex Problems

- Many problems have block structures in nature.

- We can have faster numerical convergence rates by leveraging block structures of the problem.

Motivation: Block Structured Nonconvex Problems

- Matrix Factorization [Jain et al., 2013]

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times k},\mathbf{Y}\in\mathbb{R}^{m\times k}}{\text{minimize}} \ \frac{1}{2}\|\mathbf{X}\mathbf{Y}^T - \mathbf{M}\|_F^2$$

- Matrix Sensing [Sun et al., 2014]

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times k},\mathbf{Y}\in\mathbb{R}^{m\times k}}{\text{minimize}} \ \frac{1}{2}\|\mathcal{A}(\mathbf{X}\mathbf{Y}^T - \mathbf{M})\|_F^2$$

$\mathcal{A}$: linear measurement operator and satisfies the restricted isometry property (RIP) condition

Motivation of This Work

Can we solve the nice block structured nonconvex problems to SOSP?

Alternating Gradient Descent

- Iterates of A-GD [Bertsekas 1999]:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \nabla_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) \tag{1}$$

$$\mathbf{y}^{(t+1)} = \mathbf{y}^{(t)} - \eta \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) \tag{2}$$

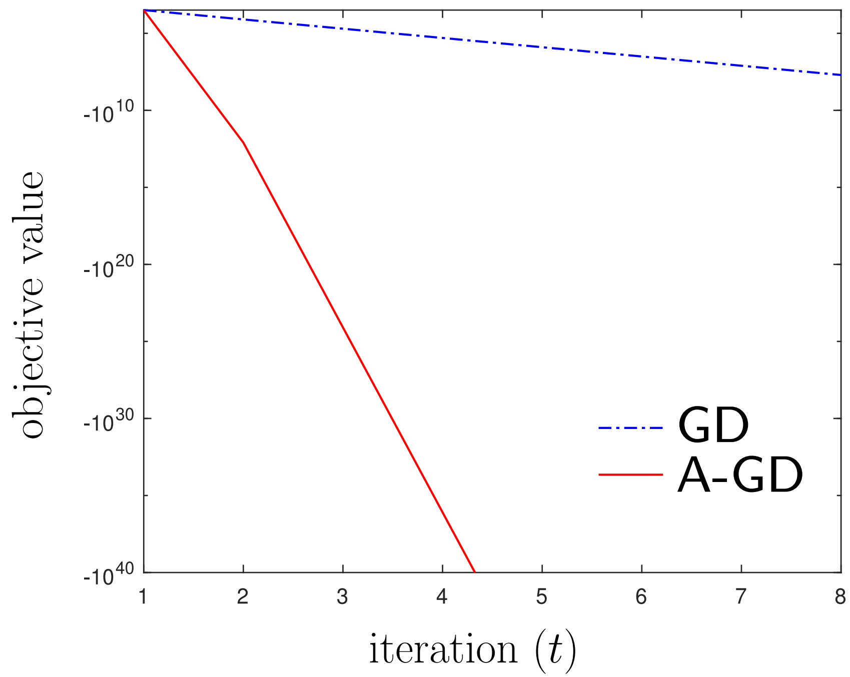
- Step-size: $\eta \leq 1/L_{\max}$

Motivation of Alternating Gradient Descent

$$\underset{x_1, x_2}{\text{minimize}} \quad \mathbf{x}^T \mathbf{M} \mathbf{x}$$

$$\mathbf{M} = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}$$

- Whole problem: $L = 1 + a$
- Block-wise: $L_{\max} = 1$



$a = 1000$

Motivation of Alernating Gradient Descent

- A-GD:
  - numerically good
  - may take a long time to escape from saddle points

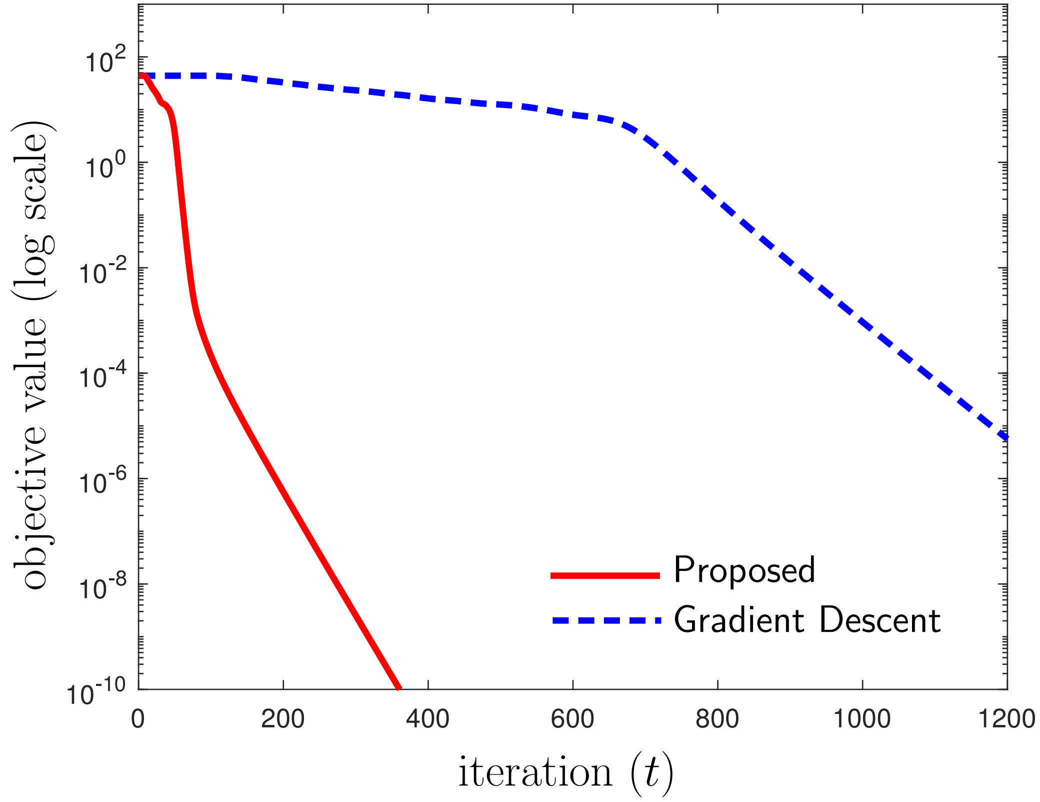
- PA-GD: numerically good and convergence rate guarantees

Matrix Factorization

A two-layer linear neural network:

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times k},\mathbf{V}\in\mathbb{R}^{m\times k}}{\text{minimize}} \sum_{i=1}^{l} \|\widehat{\boldsymbol{y}}_i - \mathbf{U}\mathbf{V}^T\widehat{\boldsymbol{x}}_i\|_2^2 = \|\widehat{\boldsymbol{Y}} - \mathbf{U}\mathbf{V}^T\widehat{\boldsymbol{X}}\|_F^2, \tag{3}$$

- $\widehat{\boldsymbol{Y}}$ and $\widehat{\boldsymbol{X}}$: $n = 100, m = 40, k = 20, l = 20$, $\mathcal{CN}(0,1)$



Convergence comparison between GD and PA-GD for learning a two-layer neural network, where $\epsilon = 10^{-10}$, $g_{\text{th}} = \epsilon/10$, $t_{\text{th}} = 10/\epsilon^{1/2}$, $r = \epsilon/10$.

Connection with Existing Works

Algorithm	Iterations	(ϵ, γ)-SOSP
PGD [Jin et al, 2017]	$\widetilde{\mathcal{O}}(1/\epsilon^2)$	$(\epsilon, \epsilon^{1/2})$
NEON+SGD [Xu and Yang, 2017]	$\widetilde{\mathcal{O}}(1/\epsilon^4)$	$(\epsilon, \epsilon^{1/2})$
NEON2+SGD [Allen-Zhu and Li, 2017]	$\widetilde{\mathcal{O}}(1/\epsilon^4)$	$(\epsilon, \epsilon^{1/2})$
NEON$^+$ [Xu et al, 2017]	$\widetilde{\mathcal{O}}(1/\epsilon^{7/4})$	$(\epsilon, \epsilon^{1/2})$
Accelerated PGD [Jin et al, 2018]	$\widetilde{\mathcal{O}}(1/\epsilon^{7/4})$	$(\epsilon, \epsilon^{1/2})$
BCD [Song et al, 2017]	N/A	$(0, 0)$
BCD [Lee et al, 2017]	N/A	$(0, 0)$
PA-GD [This Work]	$\widetilde{\mathcal{O}}(1/\epsilon^2)$	$(\epsilon, \epsilon^{1/2})$

Convergence rates of algorithms to SOSPs with the first order information, where $p \geq 4$.

Connection with Existing Works

	Asymptotic convergence to SOSPs	Convergence rate to SOSPs
Gradient descent	Lee, et al, 2017	Jin, et al, 2017
Alternating gradient descent	Lee, et al, 2017 Song, et al, 2017	This Work

Challenge of the Problem

- Variable Coupling

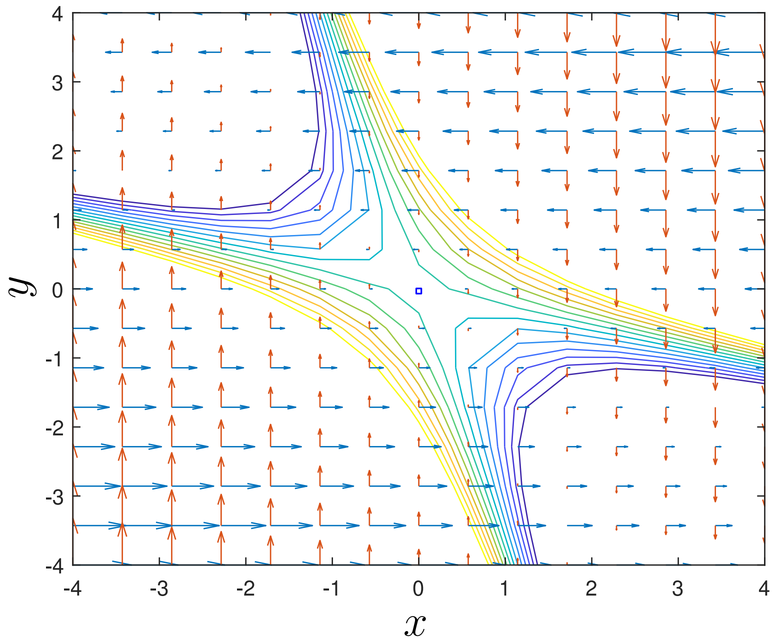
- Consider a biconvex objective function

$$f(x,y) = [x,y]\begin{bmatrix}1 & 2\\ 2 & 1\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}$$

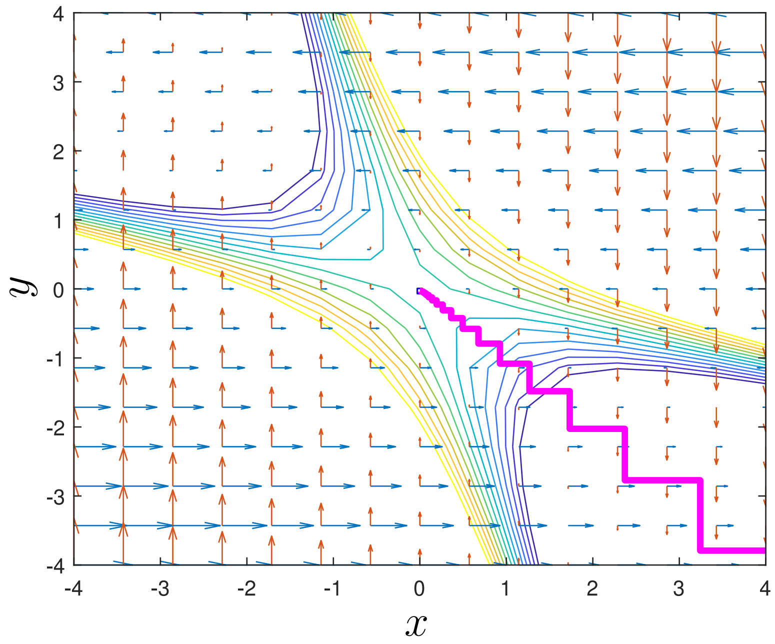
  - Block-wise: convex
  - Whole problem: nonconvex !

Adding Random Noise

- Initialize iterates at $(0, 0)$



A-GD



A-GD + random noise

Perturbed Gradient Descent

- Perturbed gradient descent [Jin, et al 2017]

For $t = 1, \ldots,$

Step 1: Gradient descent

Step 2: If the size of gradient is small (near saddle points)

Add perturbation (extract negative curvature)

Step 3: If no decrease after perturbation over t_{th} iterations

return

Perturbed Alternating Gradient Descent

Let $\mathbf{z}^{(t)} = \begin{bmatrix} \mathbf{x}^{(t)} \\ \mathbf{y}^{(t)} \end{bmatrix}$

Thresholds:

- g_{th}: gradient size
- f_{th}: objective value
- t_{th}: number of iteration

Input: $\mathbf{z}^{(1)}$, η, r, g_{th}, f_{th}, t_{th}

For $t = 1, \ldots,$

Update $\mathbf{x}^{(t+1)}$ by A-GD

If $\|\nabla_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\|^2 + \|\nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)})\|^2 \leq g_{\text{th}}^2$ and $t - t_{\text{pert}} > t_{\text{th}}$

Add random perturbation to $\mathbf{z}^{(t)}$

Update $\mathbf{x}^{(t+1)}$ by A-GD

EndIf

Update $\mathbf{y}^{(t+1)}$ by A-GD

If $t - t_{\text{pert}} = t_{\text{th}}$ and $f(\mathbf{z}^{(t)}) - f(\widetilde{\mathbf{z}}^{(t_{\text{pert}})}) > -f_{\text{th}}$

return $\widetilde{\mathbf{z}}^{(t_{\text{pert}})}$

EndIf

Perturbed Alternating Gradient Descent

- Add perturbation

  $\widetilde{\mathbf{z}}^{(t)} \leftarrow \mathbf{z}^{(t)}$ and $t_{\mathsf{pert}} \leftarrow t$

  $\mathbf{z}^{(t)} = \widetilde{\mathbf{z}}^{(t)} + \xi^{(t)}$, random noise $\xi^{(t)}$ follows uniform distribution in the interval $[0, r]$

- t_{th}: the minimum number of iterations between adding two perturbations

Main Assumptions

A1. Function $f(\mathbf{x})$: smooth and has Lipschitz continuous gradient:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \leq L\|\mathbf{x} - \mathbf{x}'\|, \ \forall \mathbf{x}, \mathbf{x}'$$

A2. Function $f(\mathbf{x})$: smooth and has block-wise Lipschitz continuous gradient:

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y})\| \leq L_{\mathbf{x}}\|\mathbf{x} - \mathbf{x}'\|, \ \forall \mathbf{x}, \mathbf{x}'$$
$$\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}')\| \leq L_{\mathbf{y}}\|\mathbf{y} - \mathbf{y}'\|, \ \forall \mathbf{y}, \mathbf{y}'.$$

Further, let $L_{\max} := \max\{L_{\mathbf{x}}, L_{\mathbf{y}}\} \leq L$.

A3. Function $f(\mathbf{x})$ has Lipschitz continuous Hessian

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{x}')\| \leq \rho\|\mathbf{x} - \mathbf{x}'\|, \ \forall \ \mathbf{x}, \mathbf{x}'$$

Convergence Rate

Theorem 1. *Under assumptions [A1]-[A3], when step-size* $\eta \leq 1/L_{\max}$*, with high probability the iterates generated by PA-GD converge to an* ϵ*-SOSP* $(\mathbf{x}, \mathbf{y})$ *satisfying*

$$\|\nabla f(\mathbf{x}, \mathbf{y})\| \leq \epsilon, \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\mathbf{x}, \mathbf{y})) \geq -\sqrt{\rho\epsilon}$$

in the following number of iterations:

$$\widetilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\right) \tag{4}$$

where $\widetilde{\mathcal{O}}$ *hides factor polylog(*d*).*

Convergence Analysis is Challenging (One Block)

- W.L.O.G set $\mathbf{x}^{(1)} = 0$

- The recursion of gradient descent (Mean Value Theorem):

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \nabla_{\mathbf{x}} f(\mathbf{x}^{(t)}) \quad (5)$$

$$= \mathbf{x}^{(t)} - \eta \nabla_{\mathbf{x}} f(0) - \eta \left(\int_0^1 \nabla^2 f(\theta \mathbf{x}^{(t)}) d\theta \right) \mathbf{x}^{(t)} \quad (6)$$

where $\theta \in [0, 1]$

Convergence Analysis is Challenging (Two Blocks)

- Recall: $\mathbf{z}^{(t)} := \begin{bmatrix} \mathbf{x}^{(t)} \\ \mathbf{y}^{(t)} \end{bmatrix}$ and W.L.O.G set $\mathbf{z}^{(1)} = 0$
- The recursion of A-GD (Mean Value Theorem):

$$\mathbf{z}^{(t+1)} = \begin{bmatrix} \mathbf{x}^{(t+1)} \\ \mathbf{y}^{(t+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(t)} \\ \mathbf{y}^{(t)} \end{bmatrix} - \eta \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) \\ \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) \end{bmatrix} \tag{7}$$

$$= \mathbf{z}^{(t)} - \eta \nabla f(0) - \eta \int_0^1 \mathbf{H}_l^{(t)} d\theta \mathbf{z}^{(t+1)} - \eta \int_0^1 \mathbf{H}_u^{(t)} d\theta \mathbf{z}^{(t)} \tag{8}$$

where

$$\theta \in [0, 1]$$

$$\mathbf{H}_l^{(t)} := \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \nabla^2_{\mathbf{xy}} f(\theta \mathbf{x}^{(t+1)}, \theta \mathbf{y}^{(t)}) & \mathbf{0} \end{bmatrix}$$

$$\mathbf{H}_u^{(t)} := \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\theta \mathbf{x}^{(t)}, \theta \mathbf{y}^{(t)}) & \nabla^2_{\mathbf{xy}} f(\theta \mathbf{x}^{(t)}, \theta \mathbf{y}^{(t)}) \\ \mathbf{0} & \nabla^2_{\mathbf{yy}} f(\theta \mathbf{x}^{(t+1)}, \theta \mathbf{y}^{(t)}) \end{bmatrix}.$$

Idea of Proof

- Let $\mathbf{z}^*$ be a strict saddle point, $\mathbf{H} = \nabla^2 f(\mathbf{z}^*)$ and $\mathbf{z}^{(1)} = 0$.

- The dynamic of the perturbed gradient descent iterates:

$$\mathbf{z}^{(t+1)} = (\mathbf{I} - \eta\mathbf{H})\mathbf{z}^{(t)} - \eta\Delta^{(t)}\mathbf{z}^{(t)} - \eta\nabla f(0) \tag{9}$$

- The dynamic of the PA-GD iterates:

$$\mathbf{z}^{(t+1)} = \mathbf{M}^{-1}\mathbf{T}\mathbf{z}^{(t)} - \eta\mathbf{M}^{-1}\Delta_u^{(t)}\mathbf{z}^{(t)} - \eta\mathbf{M}^{-1}\Delta_l^{(t)}\mathbf{z}^{(t+1)} \tag{10}$$

$$\mathbf{M} := \mathbf{I} + \eta\mathbf{H}_l, \quad \mathbf{T} := \mathbf{I} - \eta\mathbf{H}_u$$

$$\mathbf{H}_u = \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\mathbf{z}^*) & \nabla^2_{\mathbf{xy}} f(\mathbf{z}^*) \\ \mathbf{0} & \nabla^2_{\mathbf{yy}} f(\mathbf{z}^*) \end{bmatrix} \qquad \mathbf{H}_l = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \nabla^2_{\mathbf{yx}} f(\mathbf{z}^*) & \mathbf{0} \end{bmatrix}$$

Convergence Analysis

Lemma 1. *Under assumptions [A1]–[A3], let* $\mathbf{H} := \nabla^2 f(\mathbf{z}^*)$ *denote the Hessian matrix at an* ϵ*-SOSP* $\mathbf{z}$ *where* $\lambda_{\min}(\mathbf{H}) \leq -\gamma$ *and* $\gamma > 0$. *We have*

$$\lambda_{\max}(\mathbf{M}^{-1}\mathbf{T}) > 1 + \frac{\eta\gamma}{1 + L/L_{\max}} \tag{11}$$

Same Convergence Rate as GD and A-GD

Remark 1 *Under assumptions [A1]-[A3], when the step-size is small enough, with high probability the iterates generated by gradient descent converge to an* ϵ*-FOSP* $\mathbf{x}$ *satisfying*

$$\|\nabla f(\mathbf{x}, \mathbf{y})\| \leq \epsilon$$

in the following number of iterations:

$$\mathcal{O}\left(\frac{1}{\epsilon^2}\right).$$

Remark 2 Comparison between PA-GD and GD (A-GD)

- PA-GD has the same theoretical convergence rate as GD and A-GD up to some logarithmic factor.
- PA-GD can converge to SOSPs with provable convergence guarantee

Numerical Results: Two-layer Linear Neural Network

A two-layer linear neural network:

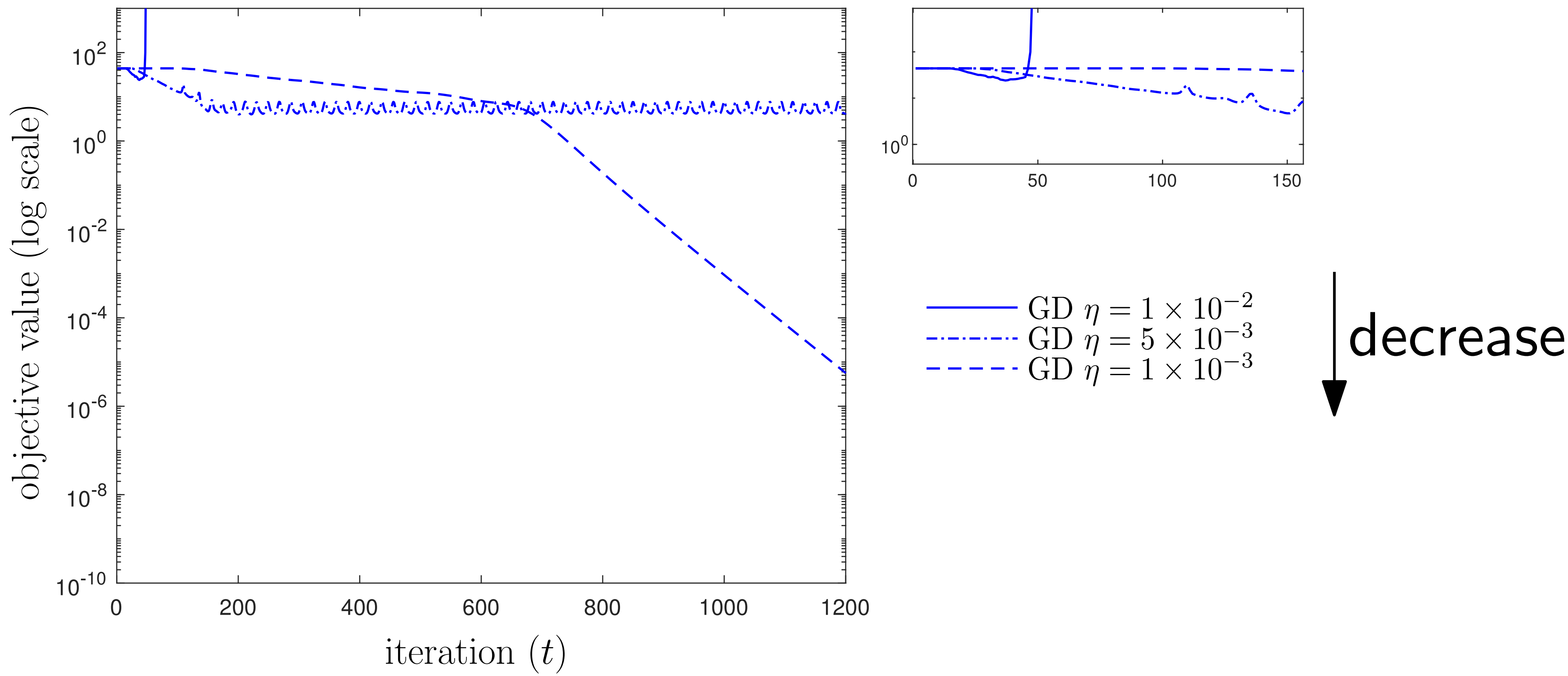
$$\underset{\mathbf{U}\in\mathbb{R}^{n\times k},\mathbf{V}\in\mathbb{R}^{m\times k}}{\text{minimize}} \sum_{i=1}^{l} \|\widehat{\boldsymbol{y}}_i - \mathbf{U}\mathbf{V}^T\widehat{\boldsymbol{x}}_i\|_2^2 = \|\widehat{\mathbf{Y}} - \mathbf{U}\mathbf{V}^T\widehat{\boldsymbol{X}}\|_F^2, \qquad (12)$$

$\widehat{\boldsymbol{X}} := [\widehat{\boldsymbol{x}}_1, \ldots, \widehat{\boldsymbol{x}}_k] \in \mathbb{R}^{m\times l}$: data matrix
$\widehat{\boldsymbol{Y}} := [\widehat{\boldsymbol{y}}_1, \ldots, \widehat{\boldsymbol{y}}_k] \in \mathbb{R}^{n\times l}$: label matrix

- $\widehat{\boldsymbol{Y}}$ and $\widehat{\boldsymbol{X}}$ are randomly generated with dimension $n = 100, m = 40, k = 20, l = 20$ and follow Gaussian distribution $\mathcal{CN}(0, 1)$
- Randomly initialize the algorithms around the origin
- Convergence comparison among GD, PGD and PA-GD for the two-layer linear neural network, where $\epsilon = 10^{-10}$, $g_{\text{th}} = \epsilon/10$, $t_{\text{th}} = 10/\epsilon^{1/2}$, $r = \epsilon/10$.

Numerical Results: Two-layer Linear Neural Network

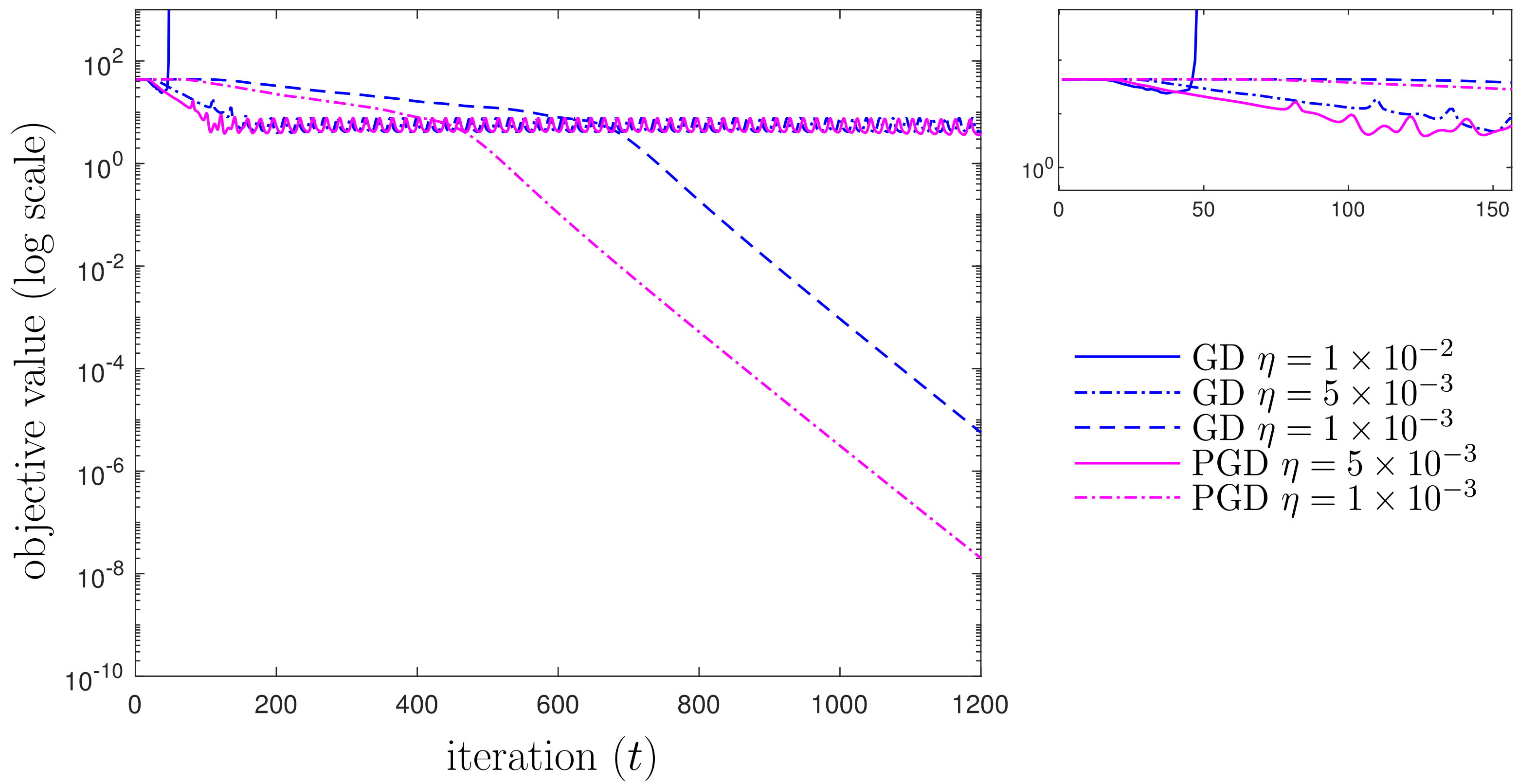
Gradient Descent:



- Different step-sizes are used to show the best GD can achieve.

Numerical Results: Two-layer Linear Neural Network

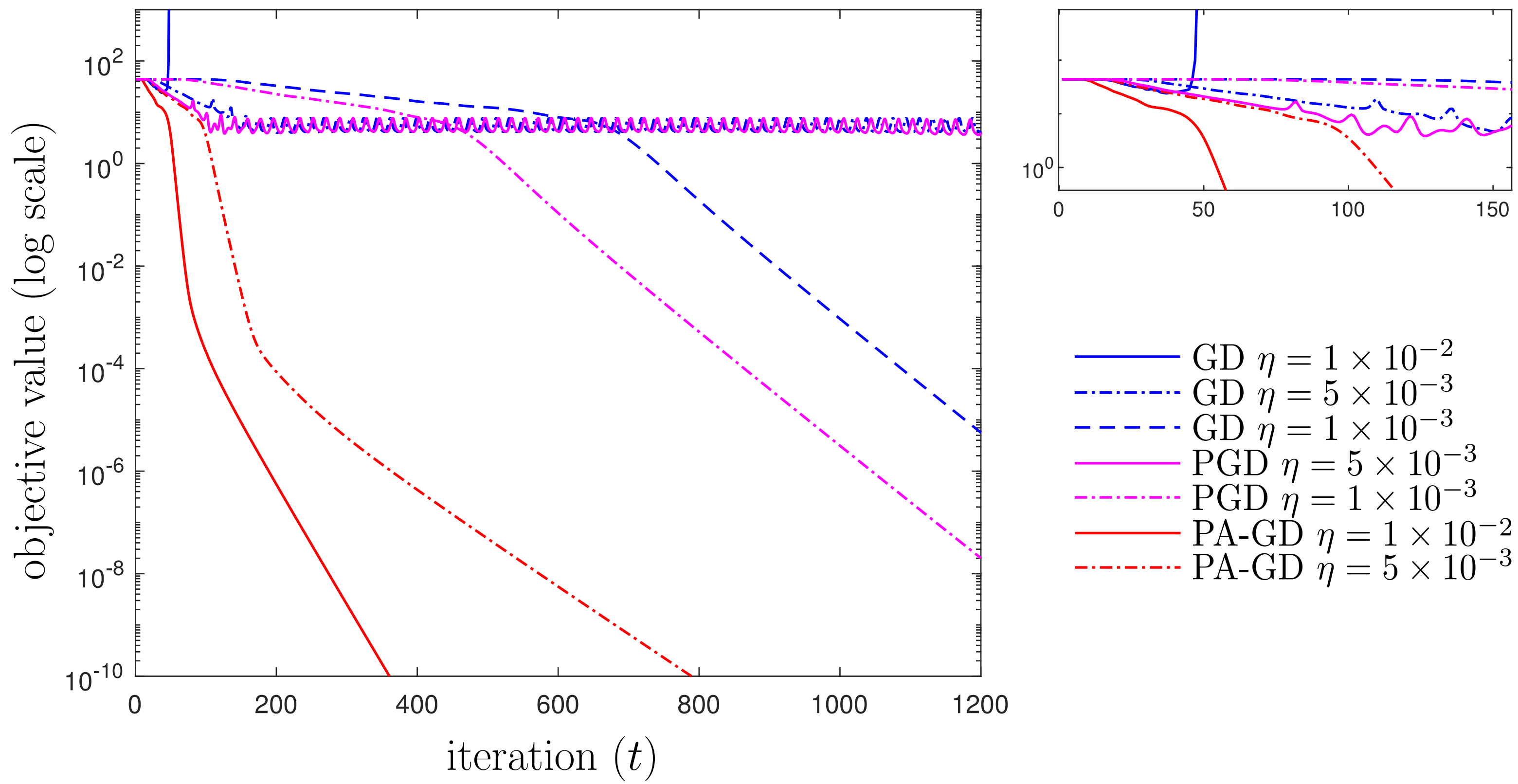
Perturbed Gradient Descent:



- The same size-sizes used in PGD

Numerical Results: Two-layer Linear Neural Network

Perturbed Alternating Gradient Descent:



- The same size-sizes used in PA-GD

Numerical Results: Matrix Factorization

Consider the matrix factorization problem as the following [Zhu, et al.' 17]:

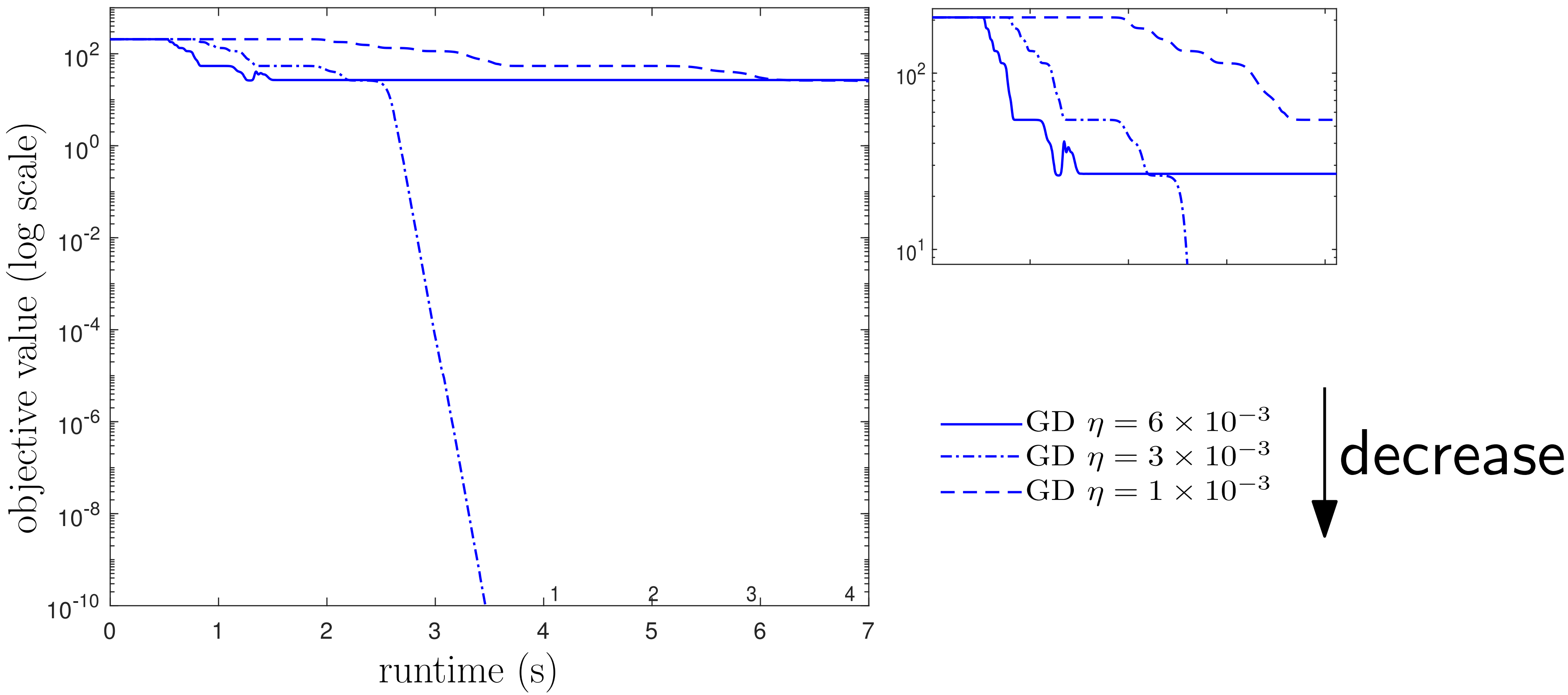
$$\underset{\mathbf{X}\in\mathbb{R}^{n\times k},\mathbf{Y}\in\mathbb{R}^{m\times k}}{\text{minimize}} \ \frac{1}{2}\|\mathbf{X}\mathbf{Y}^T - \mathbf{M}^*\|_F^2 + \frac{\mu}{4}\|\mathbf{X}^T\mathbf{X} - \mathbf{Y}^T\mathbf{Y}\|_F^2$$

where $\mu > 0$.

- Ground truth: randomly generated matrix $\mathbf{M}^* = \mathbf{U}^*(\mathbf{V}^*)^T$ with dimension $n = 200, m = 20, k = 10$
- Randomly initialize the algorithms around the origin
- Convergence comparison among GD, PGD and PA-GD for asymmetric matrix factorization, where $\epsilon = 10^{-10}$, $g_{\text{th}} = \epsilon/10$, $t_{\text{th}} = 10/\epsilon^{1/2}$, $r = \epsilon/10$.

Numerical Results: Matrix Factorization

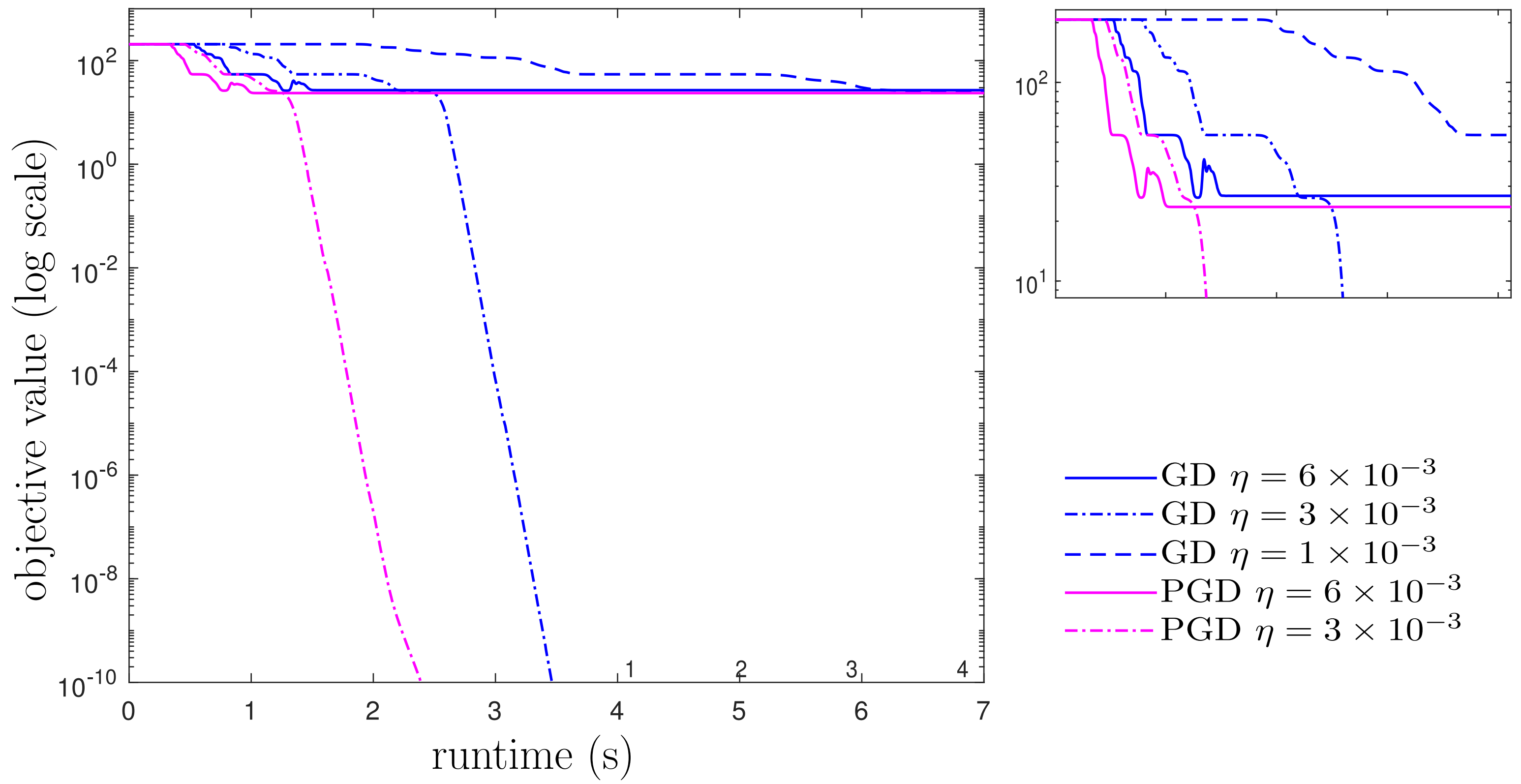
Gradient Descent:



- Different step-sizes are used to show the best GD can achieve.

Numerical Results: Matrix Factorization

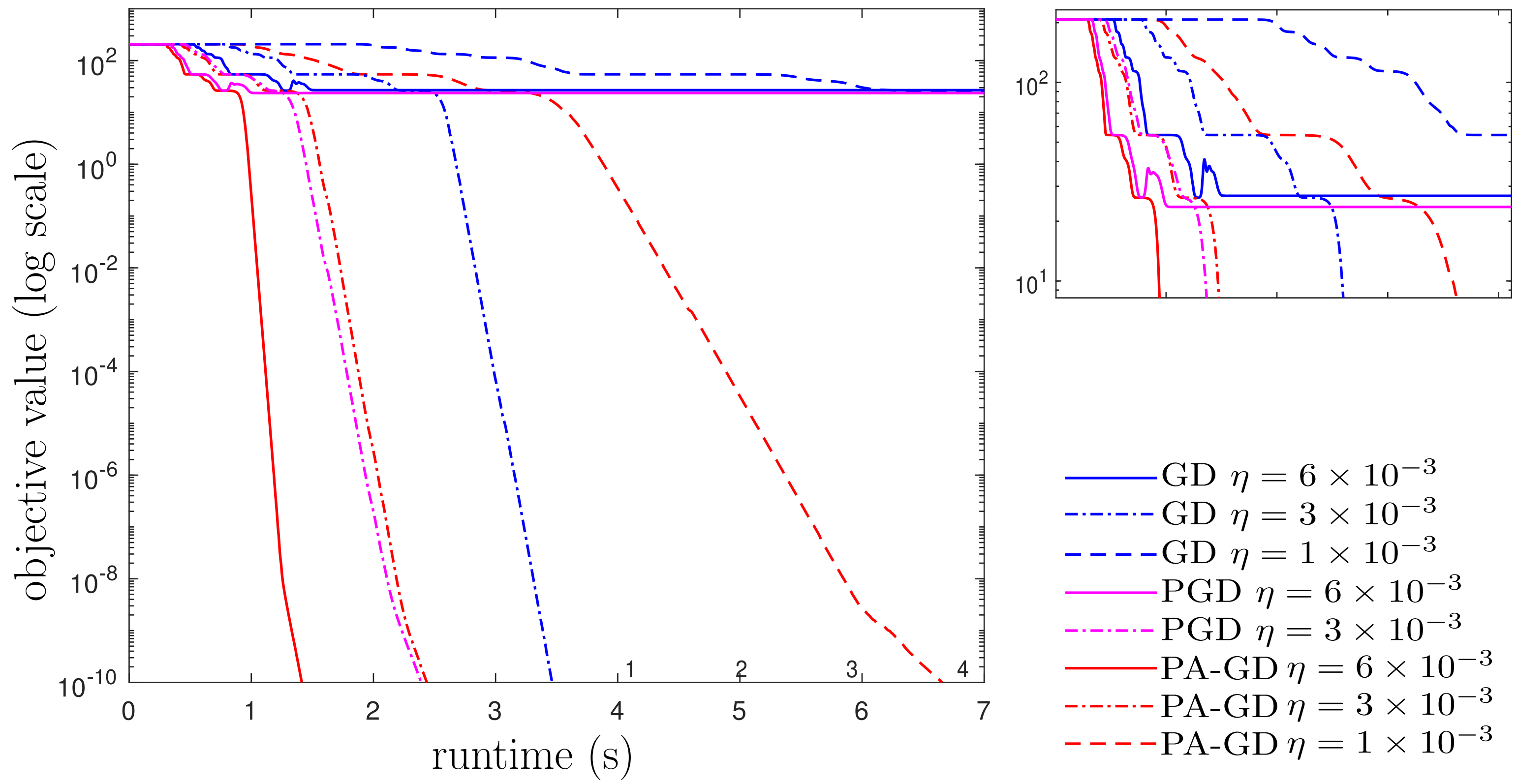
Perturbed Gradient Descent:



- The same size-sizes used in PGD

Numerical Results: Matrix Factorization

Perturbed Alternating Gradient Descent:



- The same size-sizes used in PA-GD

Conclusion, Ongoing Work and Open Problems

Conclusion:

- We consider block structured nonconvex problems:

$$\underset{\mathbf{x},\mathbf{y}}{\text{minimize}} \quad f(\mathbf{x},\mathbf{y})$$

- Convergence rate of PA-GD to SOSPs

Ongoing work:

- We consider nonconvex optimization problems with general linear inequality constraints
- Convergence rate of algorithms to SOSPs

Open Problems:

- Convergence rate of multiple blocks of coordinate descent algorithms (both unconstrained and constrained cases)

Thank You!