Fast Multi-level Optimisation Algorithms for Large Scale Machine Learning

Vahan Hovhannisyan, Panos Parpas, Stefanos Zafeiriou

SIAM Conference on Imaging Science 2016

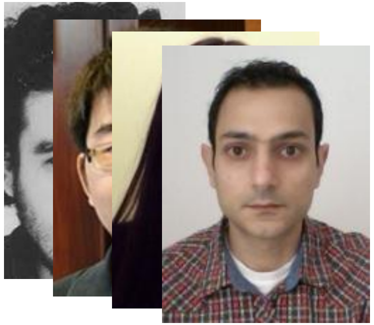
Imperial College London

Overview

- ➢ Motivation
- ➢ State of the Art
- ➢ MAGMA: A Multi-level Algorithm
- ➢ Theory and Experiments
- ➢ Discussion

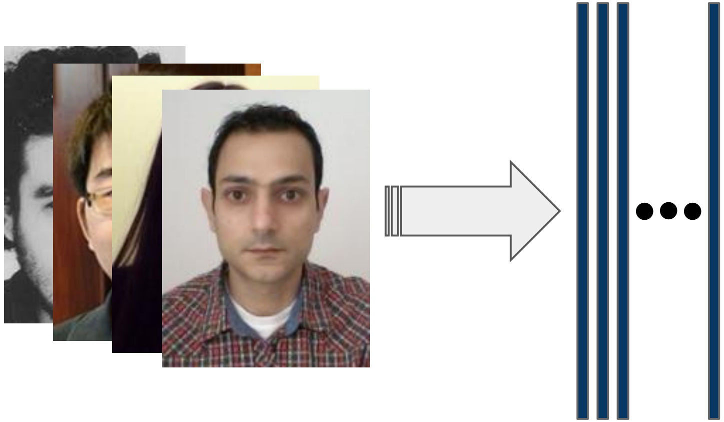
Face Recognition

Face Recognition



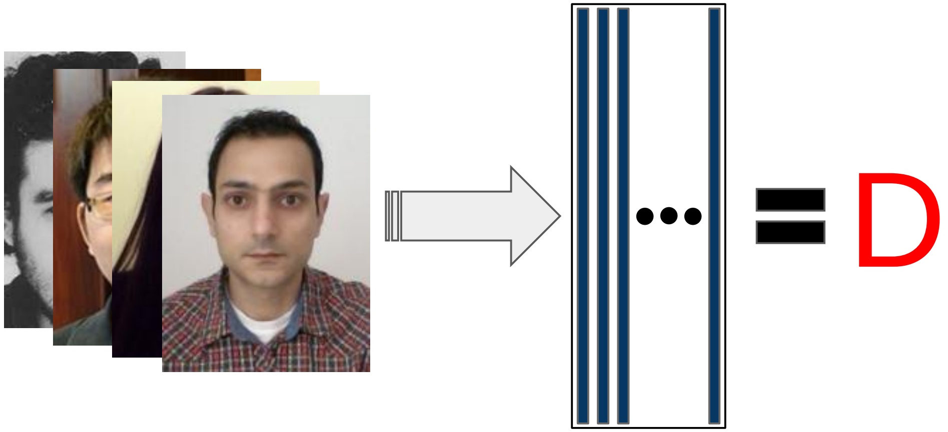
Face Recognition

Stack each image as a column vector



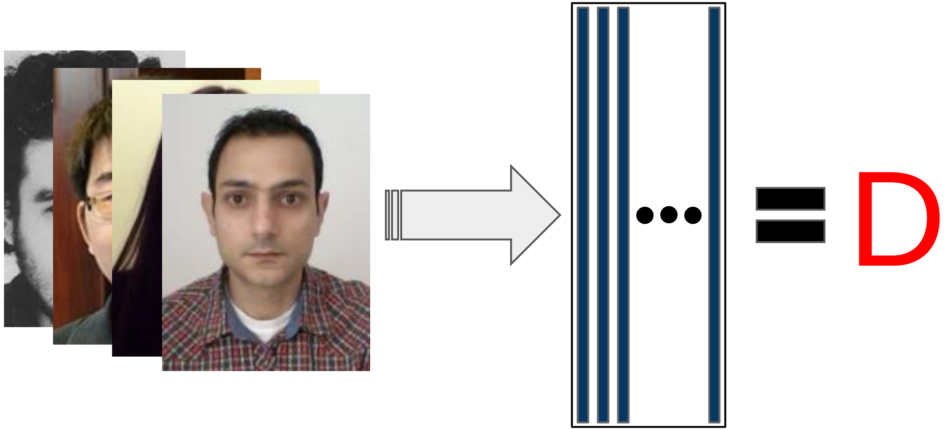
Face Recognition

Stack each image as a column vector

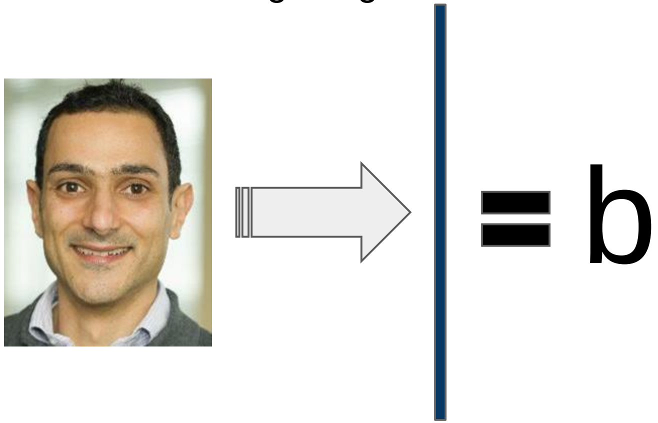


Face Recognition

Stack each image as a column vector

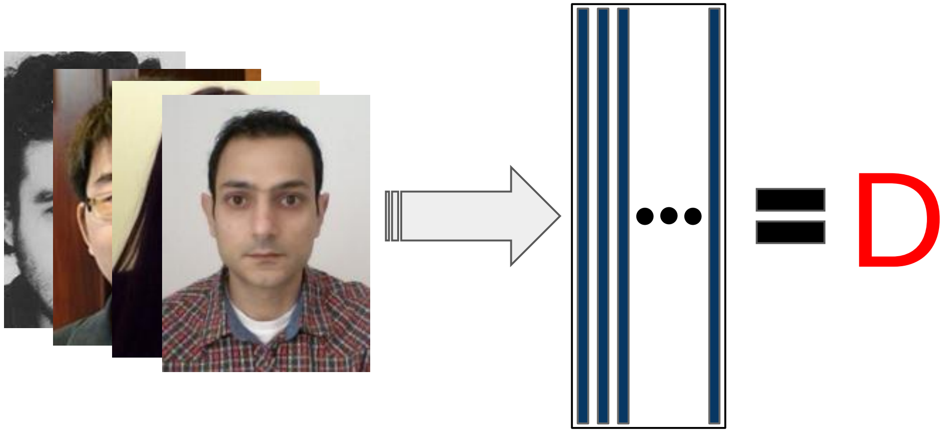


A new incoming image

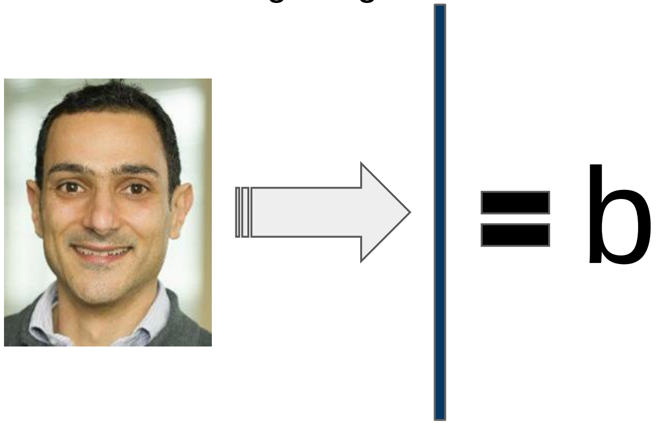


Face Recognition

Stack each image as a column vector



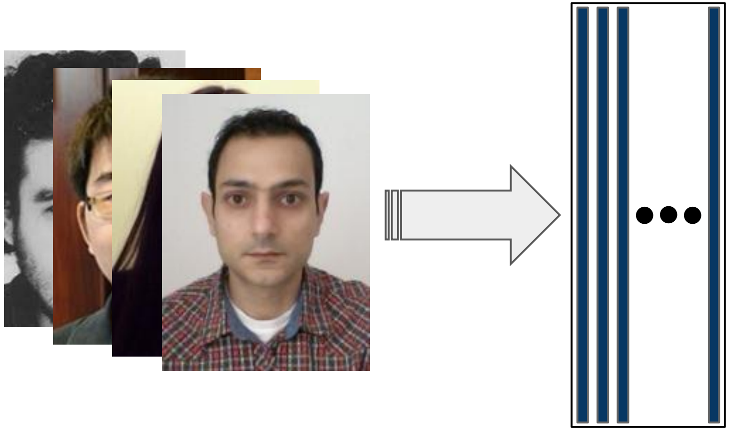
A new incoming image



Decompose **b** into **Dx** with *sparse* **x**

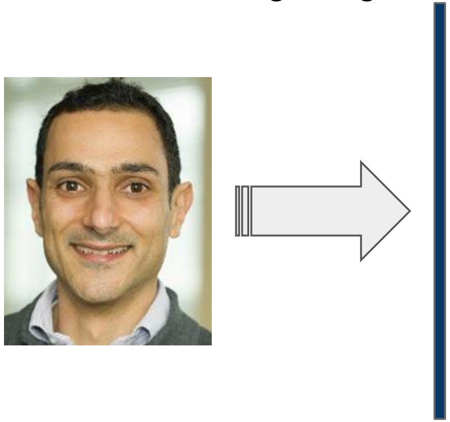
Face Recognition

Stack each image as a column vector



= D

A new incoming image

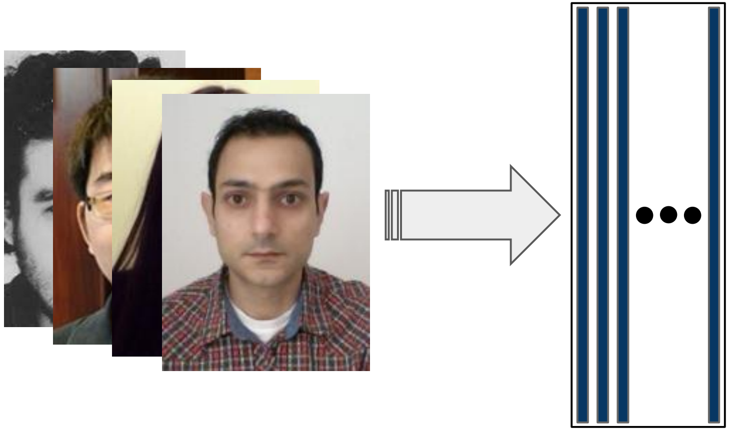


= b

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

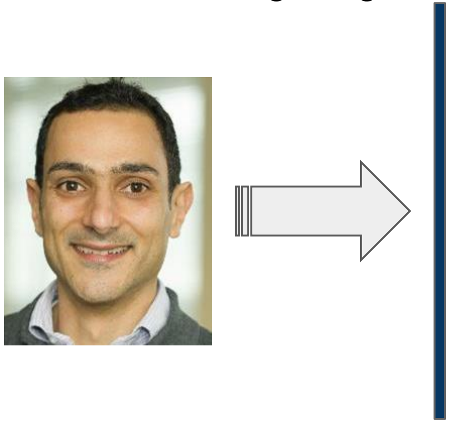
Face Recognition

Stack each image as a column vector



$= \mathbf{D}$

A new incoming image



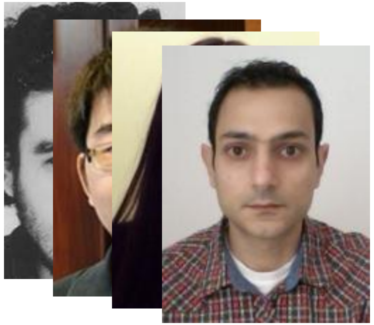
$= \mathbf{b}$

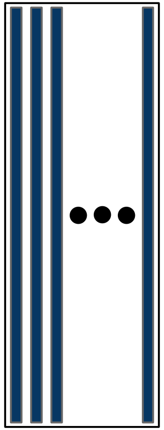
$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

LASSO

Face Recognition

Stack each image as a column vector





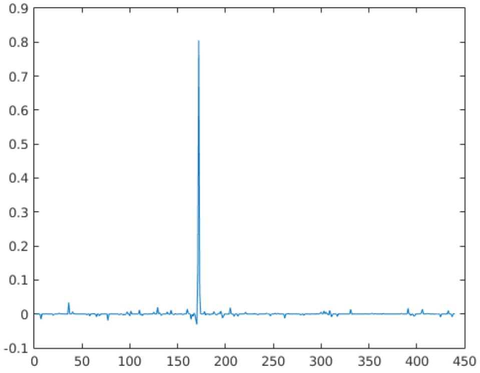
$= \mathbf{D}$

A new incoming image

$= \mathbf{b}$

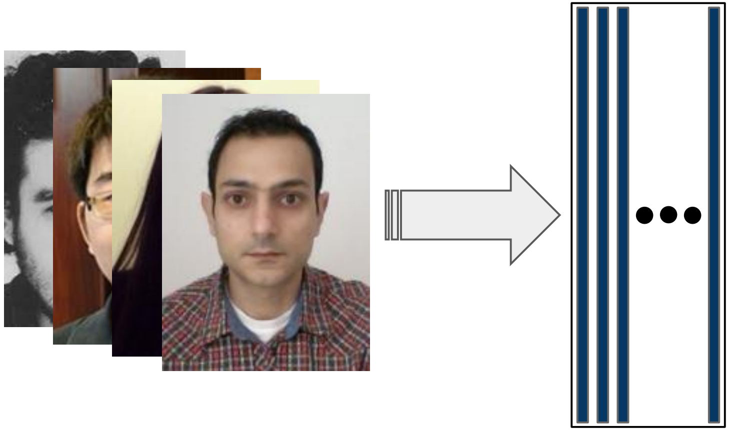
$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

LASSO



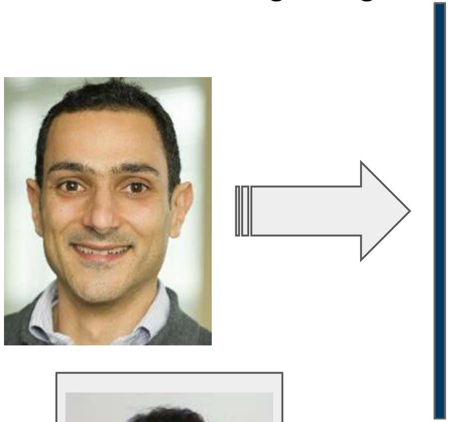
Face Recognition

Stack each image as a column vector



$= \mathbf{D}$

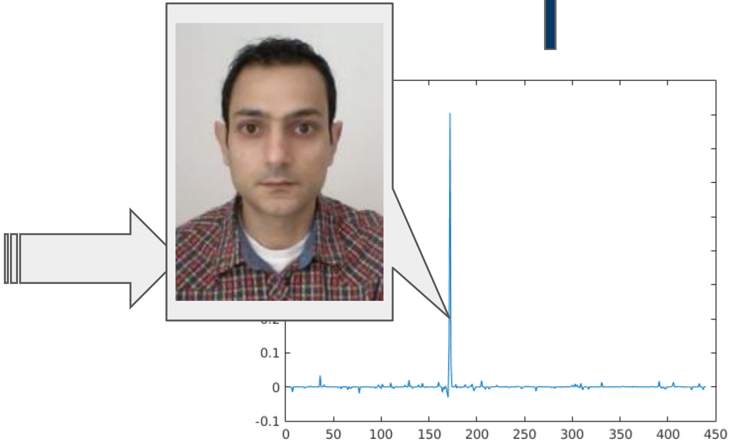
A new incoming image



$= \mathbf{b}$

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

LASSO



In general: minimize $F(x; D) = f(x; D) + g(x)$

$$\frac{1}{2}\|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2$$

$$\lambda\|\mathbf{x}\|_1$$

- f is convex and smooth with L-Lipschitz continuous gradient
- g is convex but not smooth; g is *simple*
- D is highly correlated data

Proximal Gradient

- Minimize $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, but $g(\mathbf{x})$ is not differentiable!

Proximal Gradient

- Minimize $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, but $g(\mathbf{x})$ is not differentiable!
- Instead repeatedly minimize a quadratic approximation of $F(\mathbf{x})$

Proximal Gradient

- Minimize $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, but $g(\mathbf{x})$ is not differentiable!
- Instead repeatedly minimize a quadratic approximation of $F(\mathbf{x})$
- Let $\mathcal{Q}(\mathbf{x}, \mathbf{y}; \alpha) := f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) \rangle + \frac{1}{2\alpha} \|\mathbf{x} - \mathbf{y}\|^2 + g(\mathbf{x})$

Proximal Gradient

- Minimize $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, but $g(\mathbf{x})$ is not differentiable!
- Instead repeatedly minimize a quadratic approximation of $F(\mathbf{x})$
- Let $\mathcal{Q}(\mathbf{x}, \mathbf{y}; \alpha) := f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) \rangle + \frac{1}{2\alpha} \|\mathbf{x} - \mathbf{y}\|^2 + g(\mathbf{x})$
- $\mathrm{p}(\mathbf{y}; \alpha) := \mathrm{argmin}_{\mathbf{x}} \{\mathcal{Q}(\mathbf{x}, \mathbf{y}; \alpha)\}$ is given in closed form

Proximal Gradient

- Minimize $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, but $g(\mathbf{x})$ is not differentiable!
- Instead repeatedly minimize a quadratic approximation of $F(\mathbf{x})$
- Let $\mathcal{Q}(\mathbf{x}, \mathbf{y}; \alpha) := f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) \rangle + \frac{1}{2\alpha} \|\mathbf{x} - \mathbf{y}\|^2 + g(\mathbf{x})$
- $\mathrm{p}(\mathbf{y}; \alpha) := \mathrm{argmin}_{\mathbf{x}} \{\mathcal{Q}(\mathbf{x}, \mathbf{y}; \alpha)\}$ is given in closed form

Repeatedly apply $\mathbf{x} = \mathrm{p}(\mathbf{x}; 1/L)$ until convergence

Algorithms

2 main approaches

Algorithms

2 main approaches

- Full gradient methods

Algorithms

2 main approaches

- ➢ Full gradient methods
  - ○ Use all *D* and all *x* for each update

Algorithms

2 main approaches

- Full gradient methods
  - Use all D and all x for each update
  - Impractically slow when D is large

Algorithms

2 main approaches

- Full gradient methods
  - Use all *D* and all *x* for each update
  - Impractically slow when *D* is large
- Coordinate gradient methods

Algorithms

2 main approaches

- Full gradient methods
  - Use all *D* and all *x* for each update
  - Impractically slow when *D* is large
- Coordinate gradient methods
  - Use parts of *x* and *D* for each update

Algorithms

2 main approaches

- ➢ Full gradient methods
  - Use all *D* and all *x* for each update
  - Impractically slow when *D* is large
- ➢ Coordinate gradient methods
  - Use parts of *x* and *D* for each update
  - Can be faster, but *not appropriate* when *D* has highly correlated columns

By exploiting the data structure we can make the problem smaller while using all the data.

Accelerated Proximal Gradient*

*Nesterov, Y. Smooth minimization of non-smooth functions. Mathematical programming, 2005.

Accelerated Proximal Gradient*

1. Set y to minimize quadratic approximation of F around x

*Nesterov, Y. Smooth minimization of non-smooth functions. Mathematical programming, 2005.

Accelerated Proximal Gradient*

1. Set y to minimize quadratic approximation of F around x
2. Choose step-sizes and acceleration parameters t.

*Nesterov, Y. Smooth minimization of non-smooth functions. Mathematical programming, 2005.

Accelerated Proximal Gradient*

1. Set y to minimize quadratic approximation of F around x
2. Choose step-sizes and acceleration parameters t.
3. Set z to minimize quadratic approximation of F around z

*Nesterov, Y. Smooth minimization of non-smooth functions. Mathematical programming, 2005.

Accelerated Proximal Gradient*

1. Set *y* to minimize quadratic approximation of *F* around *x*
2. Choose step-sizes and acceleration parameters t.
3. Set *z* to minimize quadratic approximation of *F* around *z*
4. Set *z* = t**y* + (1 - t)**z*

*Nesterov, Y. Smooth minimization of non-smooth functions. Mathematical programming, 2005.

Accelerated Proximal Gradient*

1. Set *y* to minimize quadratic approximation of *F* around *x*
2. Choose step-sizes and acceleration parameters t.
3. Set *z* to minimize quadratic approximation of *F* around *z*
4. Set *z* = t**y* + (1 - t)**z*

Theorem. After *T* iterations of APG it holds that

$$F(\mathbf{y}) - F(\mathbf{x}^\star) \leq \mathcal{O}\left(\frac{L}{T^2}\right)$$

*Nesterov, Y. Smooth minimization of non-smooth functions. Mathematical programming, 2005.

Accelerated Proximal Gradient*

coarse, if useful

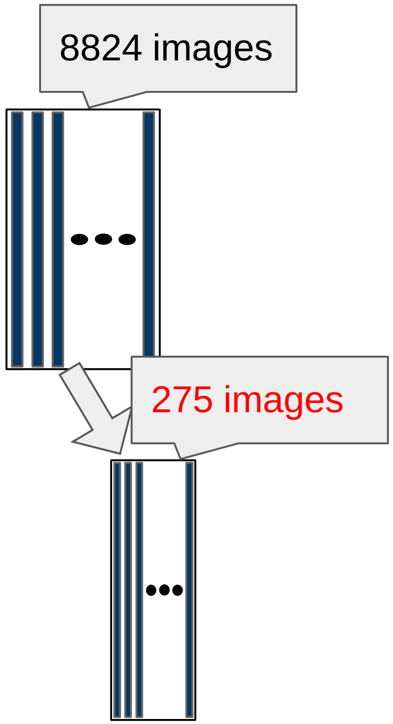
1. Set *y* to minimize quadratic approximation of *F* around *x*
2. Choose step-sizes and acceleration parameters t.
3. Set *z* to minimize quadratic approximation of *F* around *z*
4. Set z = t*y + (1 - t)*z

*Nesterov, Y. Smooth minimization of non-smooth functions. Mathematical programming, 2005.

Creating Coarse Models

Creating Coarse Models

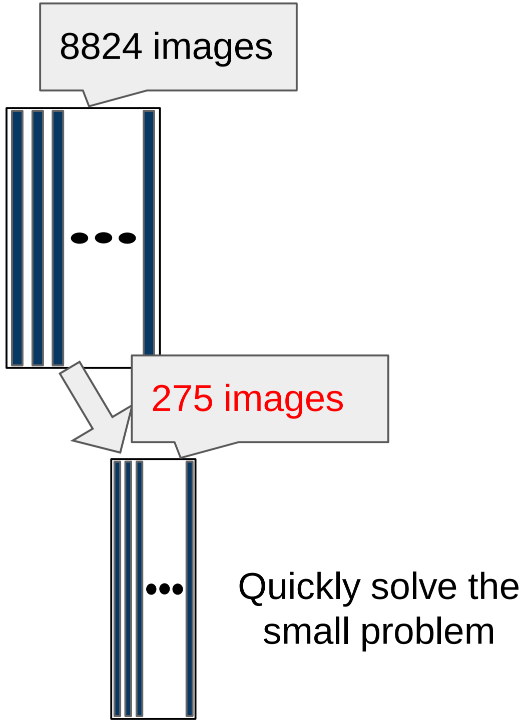
1. Create a coarse model



Creating Coarse Models

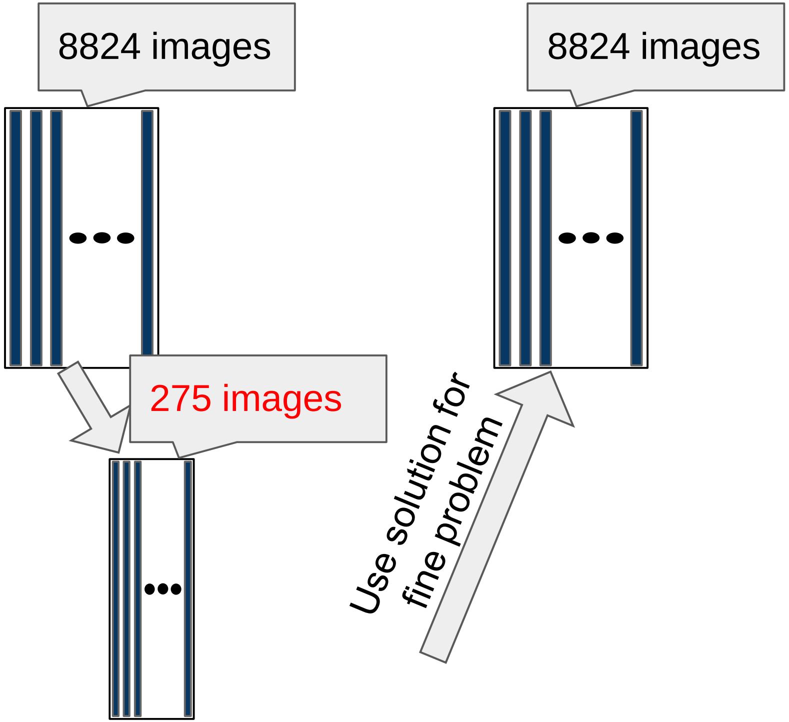
1. Create a coarse model

2. Solve the coarse model



Creating Coarse Models

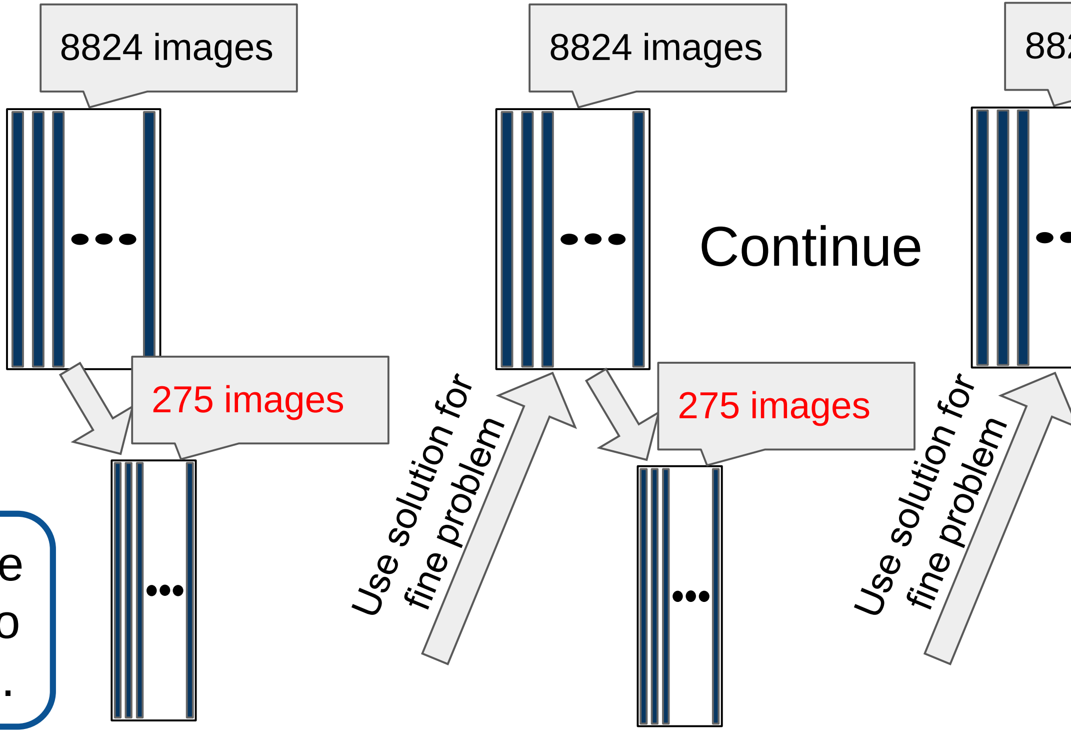
1. Create a coarse model
2. Solve the coarse model
3. Use the coarse model to iterate on the fine problem



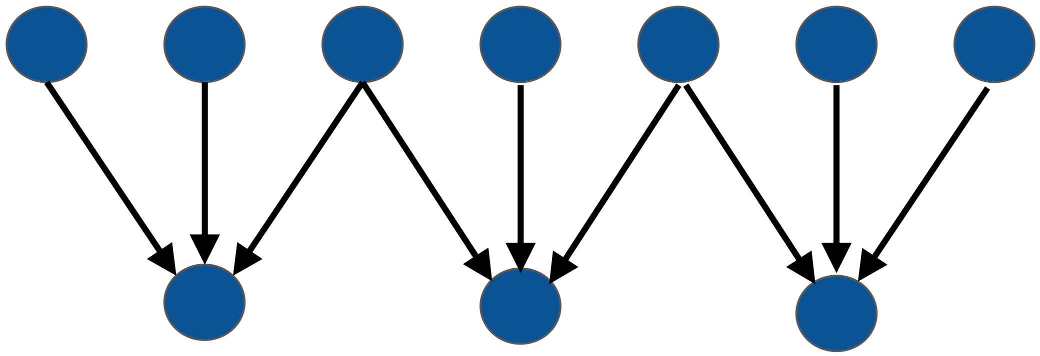
Creating Coarse Models

1. Create a coarse model
2. Solve the coarse model
3. Use the coarse model to iterate on the fine problem

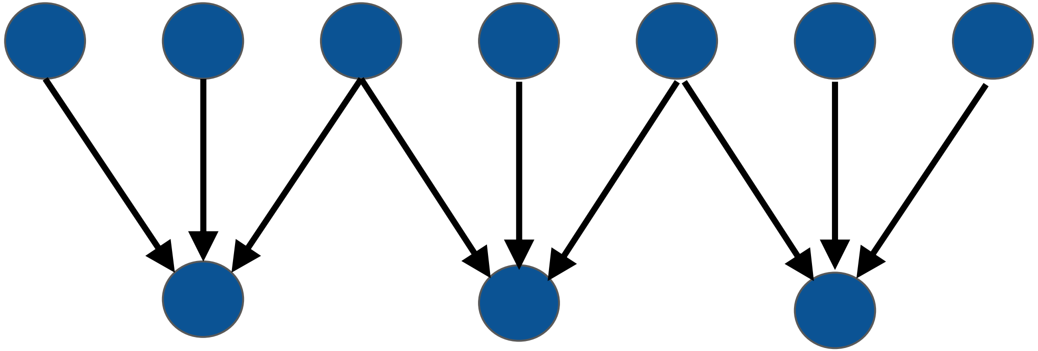
Use theory to construct appropriate step-sizes and other parameters to achieve *optimal* convergence rate.



Restriction and Prolongation

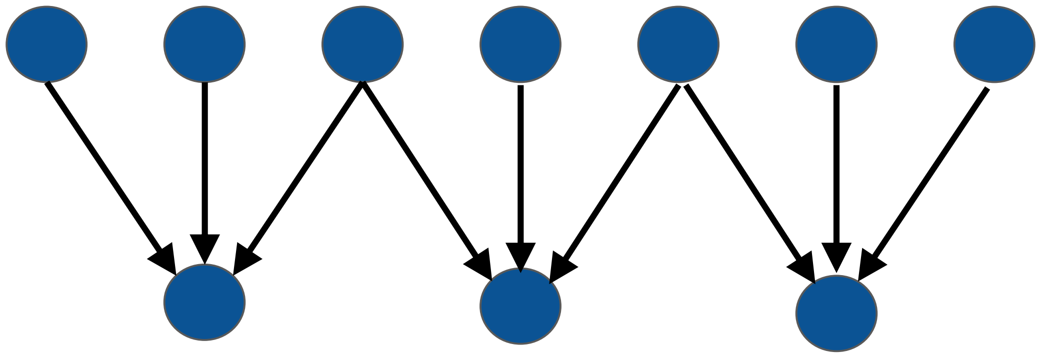


Restriction and Prolongation



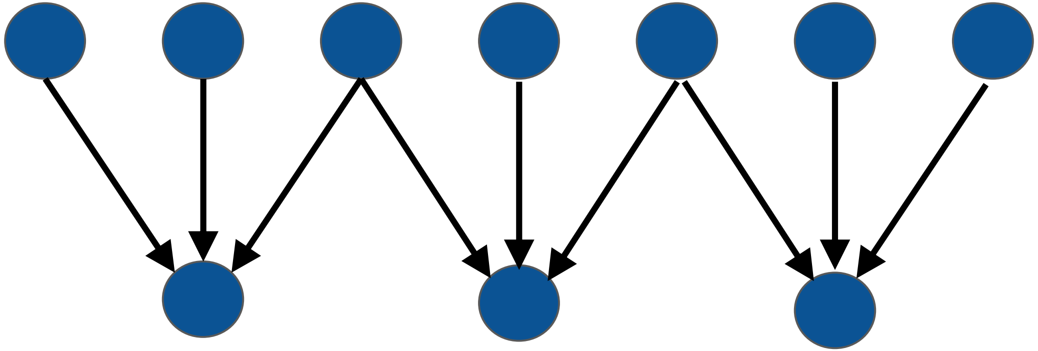
➢ $\mathbf{R} \in \mathbb{R}^{n \times \frac{n}{2}}$ linearly combines elements to half the dimension

Restriction and Prolongation



- ➢ $\mathbf{R} \in \mathbb{R}^{n \times \frac{n}{2}}$ linearly combines elements to half the dimension
- ➢ Use $\mathbf{P} = \mathbf{R}^\top \in \mathbb{R}^{\frac{n}{2} \times n}$ for prolongation

Restriction and Prolongation



- ➢ $\mathbf{R} \in \mathbb{R}^{n \times \frac{n}{2}}$ linearly combines elements to half the dimension
- ➢ Use $\mathbf{P} = \mathbf{R}^\top \in \mathbb{R}^{\frac{n}{2} \times n}$ for prolongation
- ➢ Use multiple levels to achieve smaller coarse models

Coherent Coarse Model

1. Assure that coarse and fine level problems are coherent

$$\nabla F_H(\mathbf{Rx}) = \mathbf{R}F_\mu(\mathbf{x})$$

Coherent Coarse Model

Coarse obj fun

Restriction operator

1. Assure that coarse and fine level problems are coherent

$$\nabla F_H(\mathbf{R}\mathbf{x}) = \mathbf{R}F_\mu(\mathbf{x})$$

Smooth *F*

Coherent Coarse Model

1. Assure that coarse and fine level problems are coherent

$$\nabla F_H(\mathbf{R}\mathbf{x}) = \mathbf{R}F_\mu(\mathbf{x})$$

Is achieved by adding a linear *v*-term

$$F_H(\mathbf{x}) = \boxed{f_H(\mathbf{x}) + g_H(\mathbf{x})} + \langle \mathbf{v}, \mathbf{x} \rangle$$

Coherent Coarse Model

1. Assure that coarse and fine level problems are coherent

$$\nabla F_H(\mathbf{R}\mathbf{x}) = \mathbf{R} F_\mu(\mathbf{x})$$

Is achieved by adding a linear *v*-term

$$F_H(\mathbf{x}) = \boxed{f_H(\mathbf{x}) + g_H(\mathbf{x})} + \langle \mathbf{v}, \mathbf{x} \rangle$$

where *v* is defined as

$$\mathbf{v} = \mathbf{R}\nabla F_\mu(\mathbf{x}) - (\nabla f_H(\mathbf{R}\mathbf{x}) + \nabla g_H(\mathbf{R}\mathbf{x}))$$

Coarse Correction

2. Construct a descent search direction from the coarse solution

$$d(\mathbf{x}) = \mathbf{R}^\top(\mathbf{x}_H^\star - \mathbf{R}\mathbf{x})$$

Coarse Correction

2. Construct a descent search direction from the coarse solution

$$d(\mathbf{x}) = \mathbf{R}^\top(\mathbf{x}_H^\star - \mathbf{R}\mathbf{x})$$

Sol of coarse model

Coarse Correction

2. Construct a descent search direction from the coarse solution

$$d(\mathbf{x}) = \mathbf{R}^\top (\mathbf{x}_H^\star - \mathbf{R}\mathbf{x})$$

Lemma. For any ***x*** it holds that

$$\langle d(\mathbf{x}), \nabla F_\mu(\mathbf{x}) \rangle < -\rho \|\nabla F_\mu(\mathbf{x})\|^2 \leq 0$$

Coarse Correction

2. Construct a descent search direction from the coarse solution

$$d(\mathbf{x}) = \mathbf{R}^\top(\mathbf{x}_H^\star - \mathbf{R}\mathbf{x})$$

Lemma. For any ***x*** it holds that

$$\langle d(\mathbf{x}), \nabla F_\mu(\mathbf{x})\rangle < -\rho\|\nabla F_\mu(\mathbf{x})\|^2 \leq 0$$

3. Use Armijo backtracking line search to find step-size s such that

$$F_\mu(\mathbf{y}) \leq F_\mu(\mathbf{x}) + cs\langle d(\mathbf{x}), \nabla F_\mu(\mathbf{x})\rangle$$

MAGMA

MAGMA

1. If conditions for coarse correction hold

MAGMA

1. If conditions for coarse correction hold

   choose a smoothing parameter

MAGMA

1. If conditions for coarse correction hold

   choose a smoothing parameter
   set y by a coarse correction step on x

MAGMA

1. If conditions for coarse correction hold

    choose a smoothing parameter
    set y by a coarse correction step on x

   Otherwise

MAGMA

1. If conditions for coarse correction hold

   choose a smoothing parameter
   set y by a coarse correction step on x

   Otherwise

   set y by a proximal gradient step on x

MAGMA

1. If conditions for coarse correction hold

    choose a smoothing parameter
    set y by a coarse correction step on x

   Otherwise

    set y by a proximal gradient step on x

2. Set z by a proximal gradient step on z

MAGMA

1. If conditions for coarse correction hold

   choose a smoothing parameter
   set y by a coarse correction step on x

   Otherwise

   set y by a proximal gradient step on x

2. Set z by a proximal gradient step on z

3. Update $x = t*y + (1 - t)*z$

Convergence Rate

Theorem. After T iterations of MAGMA it holds that

$$F(\mathbf{y}) - F(\mathbf{x}^\star) \leq \mathcal{O}\left(\frac{L}{T^2}\right)$$

MAGMA*

- ➢ Optimal convergence rate
- ➢ Much faster for the face recognition problem

*Hovhannisyan V, Parpas P, Zafeiriou S. MAGMA: Multi-level accelerated gradient mirror descent algorithm for large-scale convex composite minimization. Under revision at *SIAM journal on imaging sciences*.

Experiments for Face Recognition

Comparing against FISTA* and APCG**.

*Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* (2009).

**Lin Q, Lu Z, Xiao L. Accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. SIAM Journal on Optimization (2015).

Experiments for Face Recognition

Comparing against FISTA* and APCG**.

Dataset of images of *200*x*200* resolution.

*Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* (2009).

**Lin Q, Lu Z, Xiao L. Accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. SIAM Journal on Optimization (2015).

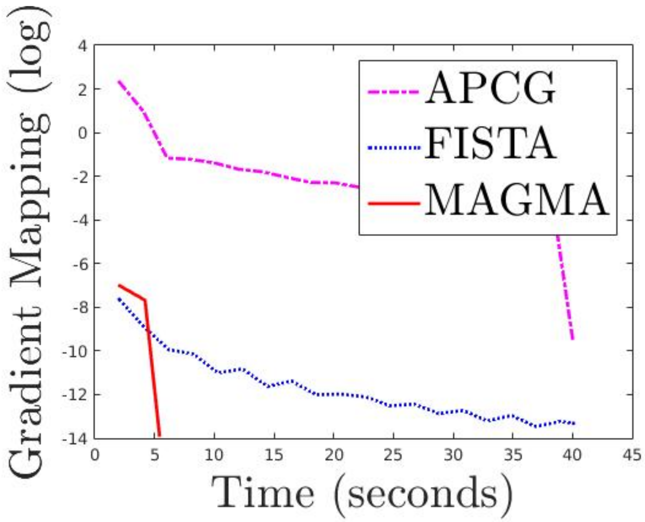
Experiments for Face Recognition

Comparing against FISTA* and APCG**.

Dataset of images of *200*x*200* resolution.

- ➢ *440* images in the dictionary

440 images in the dictionary



Time until convergence

*Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* (2009).

**Lin Q, Lu Z, Xiao L. Accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. SIAM Journal on Optimization (2015).

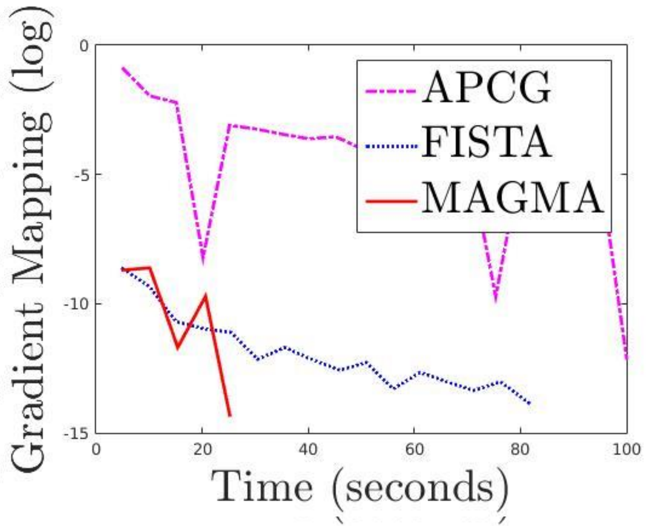
Experiments for Face Recognition

Comparing against FISTA* and APCG**.

Dataset of images of *200*x*200* resolution.

- ➢ *440* images in the dictionary
- ➢ *944* images in the dictionary

944 images in the dictionary



Time until convergence

*Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* (2009).

**Lin Q, Lu Z, Xiao L. Accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. SIAM Journal on Optimization (2015).

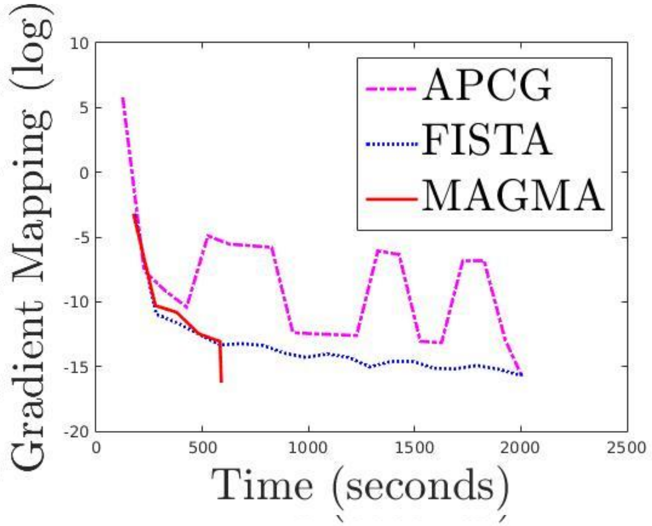
Experiments for Face Recognition

Comparing against FISTA* and APCG**.

Dataset of images of *200*x*200* resolution.

- ➢ *440* images in the dictionary
- ➢ *944* images in the dictionary
- ➢ *8824* images in the dictionary

8824 images in the dictionary



Time until convergence

*Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* (2009).

**Lin Q, Lu Z, Xiao L. Accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. SIAM Journal on Optimization (2015).

Experiments for Face Recognition

Compared against several methods, reporting only FISTA*.

Dataset of images of *200*x*200* resolution.

*Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* (2009).

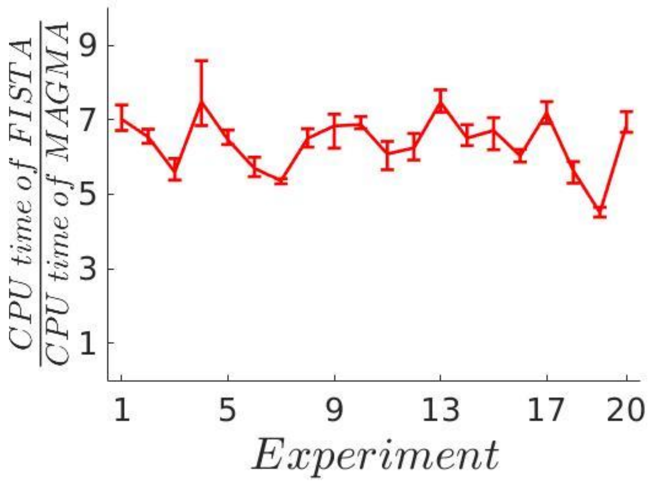
Experiments for Face Recognition

440 images in the dictionary

Compared against several methods, reporting only FISTA*.

Dataset of images of *200*x*200* resolution.

- ➢ *440* images in the dictionary



Relative CPU times for experiments with *20* different input images.

*Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* (2009).

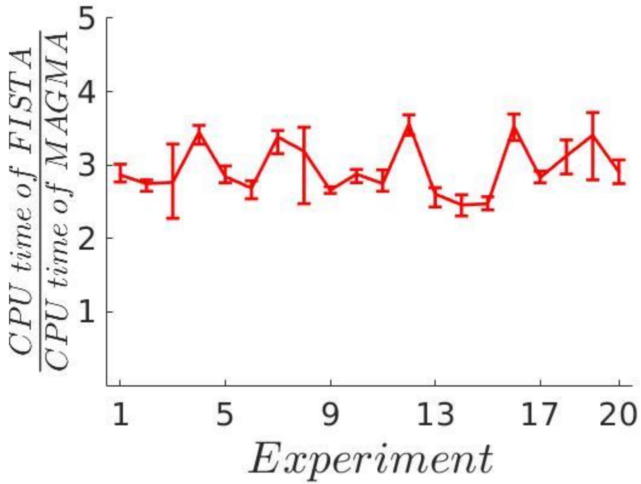
Experiments for Face Recognition

Compared against several methods, reporting only FISTA*.

Dataset of images of *200*x*200* resolution.

- ➢ *440* images in the dictionary
- ➢ *944* images in the dictionary

944 images in the dictionary



Relative CPU times for experiments with *20* different input images.

*Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* (2009).

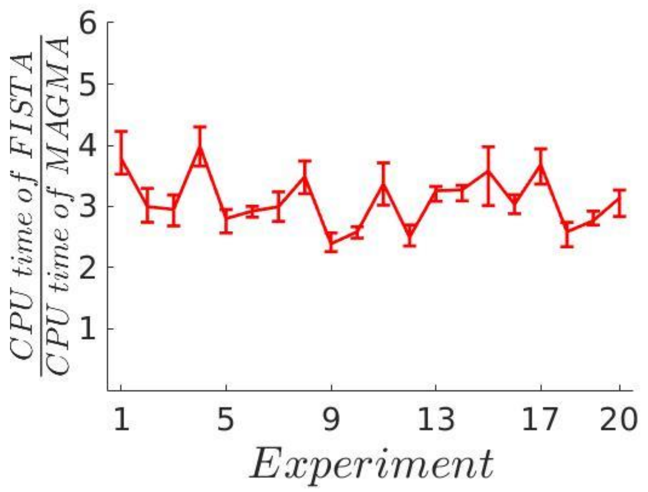
Experiments for Face Recognition

Compared against several methods, reporting only FISTA*.

Dataset of images of *200*x*200* resolution.

- ➢ *440* images in the dictionary
- ➢ *944* images in the dictionary
- ➢ *8824* images in the dictionary

8824 images in the dictionary



Relative CPU times for experiments with *20* different input images.

*Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* (2009).

Conclusion

➢ MAGMA: accelerated multi-level proximal algorithm

Conclusion

- MAGMA: accelerated multi-level proximal algorithm
- Excellent numerical results for face recognition

Conclusion

- ➢ MAGMA: accelerated multi-level proximal algorithm
- ➢ Excellent numerical results for face recognition
- ➢ Next:
  - ○ Reduce to one proximal step per iteration

Conclusion

- MAGMA: accelerated multi-level proximal algorithm
- Excellent numerical results for face recognition
- Next:
  - Reduce to one proximal step per iteration
  - Test on other problems, e.g. robust PCA and face alignment

Conclusion

- ➢ MAGMA: accelerated multi-level proximal algorithm
- ➢ Excellent numerical results for face recognition
- ➢ Next:
  - Reduce to one proximal step per iteration
  - Test on other problems, e.g. robust PCA and face alignment
  - Combine with stochastic updates