Analyzing Non-convex Optimization for Sparse Coding

Tengyu Ma
Princeton

Sanjeev Arora
Princeton

Rong Ge
Microsoft Research

Ankur Moitra
MIT

Sparse Coding (a.k.a Dictionary Learning)

- Data has sparse representation in an appropriate basis:

Sparse Coding (a.k.a Dictionary Learning)

- Data has sparse representation in an appropriate basis:

$$\begin{bmatrix} \\ y \\ \\ \end{bmatrix} \approx \begin{bmatrix} \vdots & & \vdots \\ A_1 & \cdots \quad \cdots & A_m \\ \vdots & & \vdots \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_m \end{bmatrix}$$

Sparse Coding (a.k.a Dictionary Learning)

- Data has sparse representation in an appropriate basis:

$$\begin{bmatrix} \\ y \\ \\ \end{bmatrix} \approx \begin{bmatrix} \vdots & & \vdots \\ A_1 & \cdots \quad \cdots & A_m \\ \vdots & & \vdots \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_m \end{bmatrix}$$

unknown dictionary

$(n \times m, m > n)$

Sparse Coding (a.k.a Dictionary Learning)

- Data has sparse representation in an appropriate basis:

$$\begin{bmatrix} \\ y \\ \\ \end{bmatrix} \approx \begin{bmatrix} \vdots & & & \vdots \\ A_1 & \cdots & \cdots & A_m \\ \vdots & & & \vdots \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_m \end{bmatrix}$$

unknown dictionary $(n \times m, m > n)$

representation k non-zeros

Sparse Coding (a.k.a Dictionary Learning)

- Data has sparse representation in an appropriate basis:

$$\begin{bmatrix} \\ y \\ \\ \end{bmatrix} \approx \begin{bmatrix} \vdots & & & \vdots \\ A_1 & \cdots & \cdots & A_m \\ \vdots & & & \vdots \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_m \end{bmatrix}$$

unknown dictionary $(n \times m, m > n)$

representation k non-zeros

- In compressed sensing, dict. is known and designed
- Here: learn unknown dict. (designed by nature)

Sparse Coding (a.k.a Dictionary Learning)

- Learning the basis from multiple samples

$$\begin{bmatrix} y^1 & \cdots & \cdots & y^p \end{bmatrix} \approx \begin{bmatrix} \vdots & & & \vdots \\ A_1 & \cdots & \cdots & A_m \\ \vdots & & & \vdots \end{bmatrix} \cdot \begin{bmatrix} x^1 & \cdots & \cdots & x^p \end{bmatrix}$$

unknown dictionary $(n \times m, m > n)$

representations k non-zeros each col.

Sparse Coding (a.k.a Dictionary Learning)

- Learning the basis from multiple samples

$$\begin{bmatrix} y^1 & \cdots & \cdots & y^p \end{bmatrix} \approx \begin{bmatrix} \vdots & & \vdots \\ A_1 & \cdots & A_m \\ \vdots & & \vdots \end{bmatrix} \cdot \begin{bmatrix} x^1 & \cdots & \cdots & x^p \end{bmatrix}$$

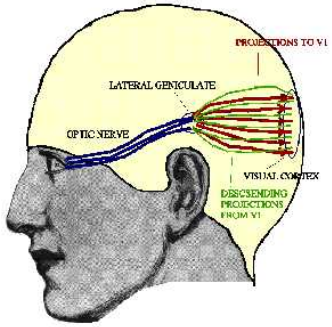
unknown dictionary $(n \times m, m > n)$

representations k non-zeros each col.

notation: $Y \approx A \cdot X$

Origins of Sparse Coding

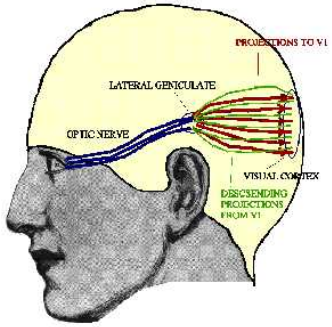
Primal visual cortex $V1$



Olshausen and Field'97:

Origins of Sparse Coding

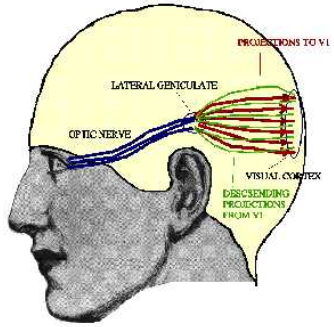
Primal visual cortex $V1$



Olshausen and Field'97:

Origins of Sparse Coding

Primal visual cortex $V1$

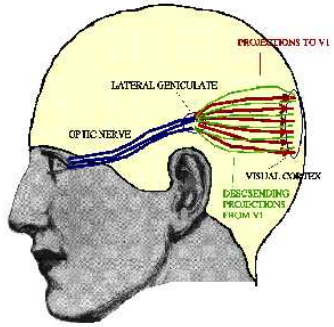


Olshausen and Field'97:

$$\rightarrow \approx \begin{bmatrix} \vdots & & \vdots \\ A_1 & \cdots & A_m \\ \vdots & & \vdots \end{bmatrix} \cdot \rightarrow \cdots\cdots$$

Origins of Sparse Coding

Primal visual cortex $V1$



Olshausen and Field'97:

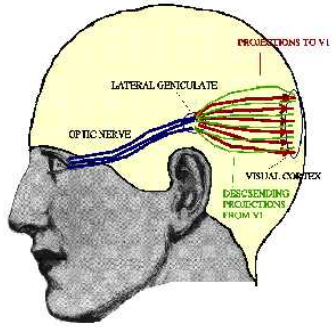
$$\rightarrow \quad \approx \begin{bmatrix} \vdots & & \vdots \\ A_1 & \cdots & A_m \\ \vdots & & \vdots \end{bmatrix} \cdot \quad \rightarrow \quad \cdots\cdots$$

neurons

neurons

Origins of Sparse Coding

Primal visual cortex $V1$



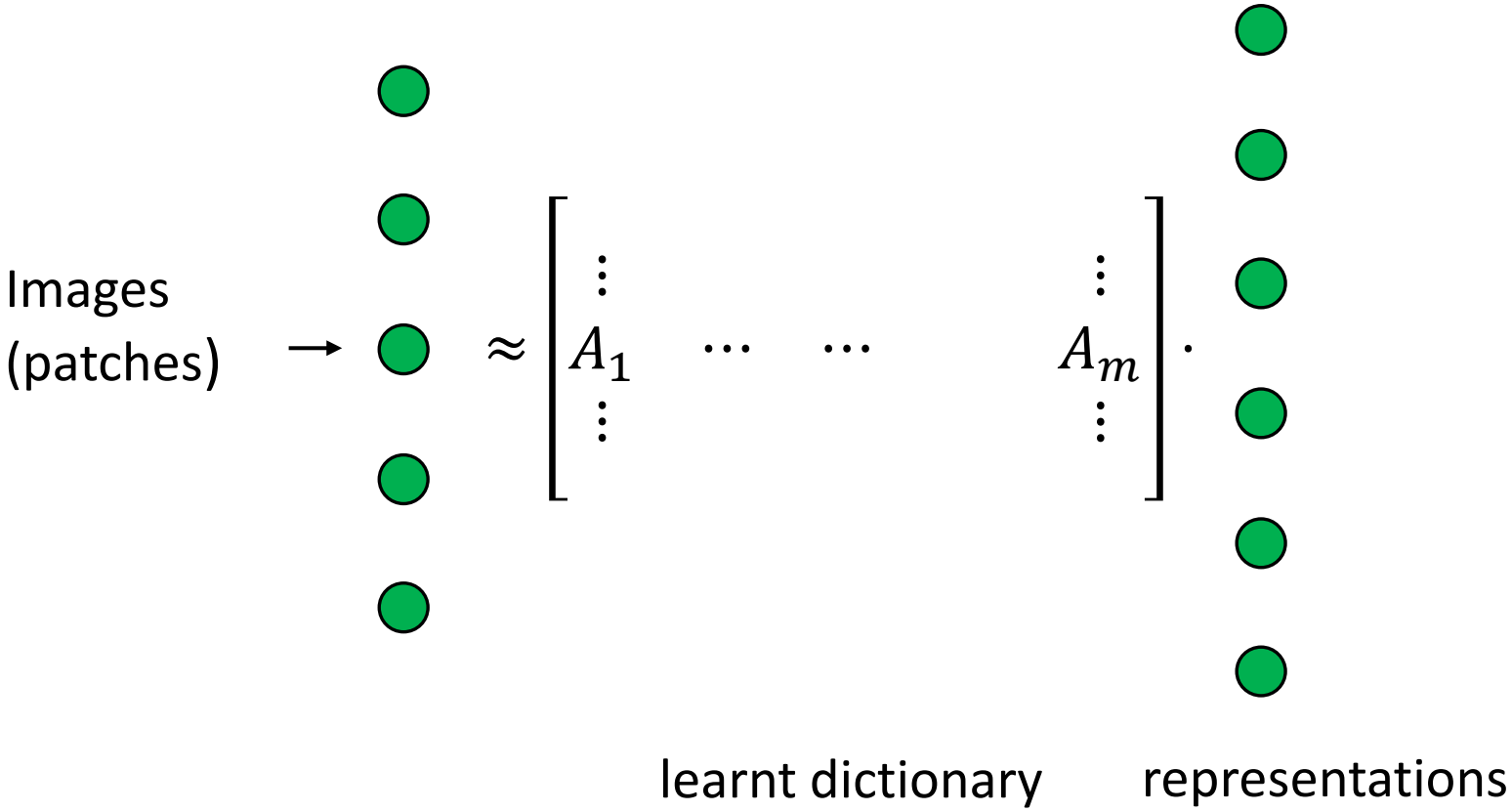
Olshausen and Field'97:

$$\rightarrow \approx \begin{bmatrix} \vdots & & \vdots \\ A_1 & \cdots \quad \cdots & A_m \\ \vdots & & \vdots \end{bmatrix} \cdot \rightarrow \cdots\cdots$$

neurons A: weights on synapses neurons

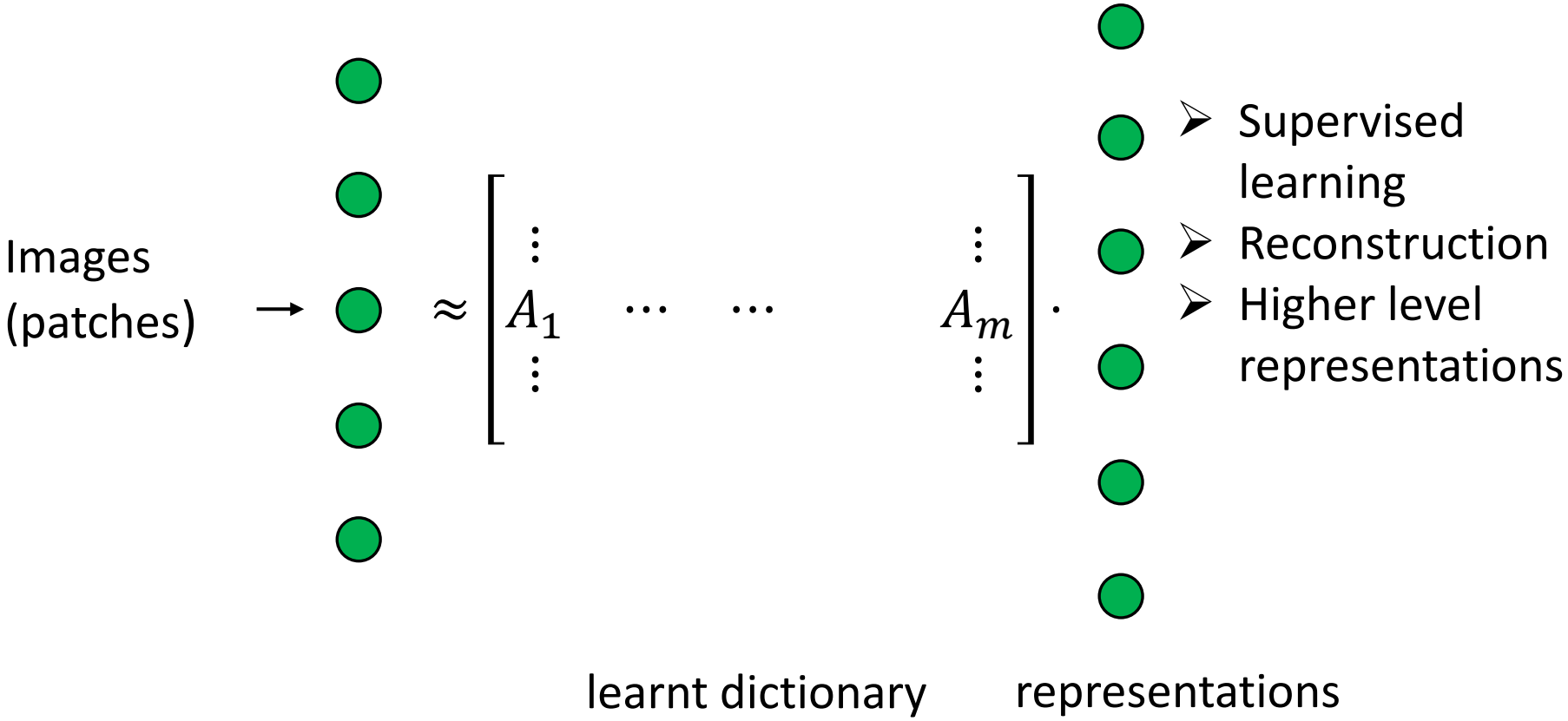
Applications to Machine Learning

- Denoising, Edge Detection, Super-resolution, Deep Learning



Applications to Machine Learning

- Denoising, Edge Detection, Super-resolution, Deep Learning



Outline

Efficient and provable algorithms for sparse coding?

- Introduction
  - Non-convex heuristics used in practice
  - Genearitve model and main results
- Analyzing alternating minimization

Non-convex Heuristics

- Recall

$$\begin{bmatrix} y^1 & \cdots & \cdots & y^p \end{bmatrix} \approx \begin{bmatrix} \vdots & & & \vdots \\ A_1 & \cdots & \cdots & A_m \\ \vdots & & & \vdots \end{bmatrix} \cdot \begin{bmatrix} x^1 & \cdots & \cdots & x^p \end{bmatrix}$$

data dictionary representation

- Energy function

$$\mathcal{E}(A, X) = \sum_{j=1}^{p} \left\| y^j - Ax^j \right\|^2 + \sum_{j=1}^{p} \ell\left(x^j\right)$$

Non-convex Heuristics

- Recall

$$\begin{bmatrix} y^1 & \cdots & \cdots & y^p \end{bmatrix} \approx \begin{bmatrix} \vdots & & & \vdots \\ A_1 & \cdots & \cdots & A_m \\ \vdots & & & \vdots \end{bmatrix} \cdot \begin{bmatrix} x^1 & \cdots & \cdots & x^p \end{bmatrix}$$

data dictionary representation

- Energy function

$$\mathcal{E}(A, X) = \sum_{j=1}^{p} \left\| y^j - Ax^j \right\|^2 + \sum_{j=1}^{p} \ell\left(x^j\right)$$

Sparse penalty, e.g.

$$\ell \cdot = \left| \cdot\cdot \quad 0 \right., \left| \cdot \quad 1 \right.$$

$$\left. 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad \right| \cdot \Big|_1$$

Minimizing Energy Function

$$\min_{A,X} \mathcal{E}(A,X) = \sum_{j=1}^{p} \left\| y^j - Ax^j \right\|^2 + \sum_{j=1}^{p} \ell\left(x^j\right)$$

Minimizing Energy Function

$$\min_{A,X} \mathcal{E}(A,X) = \sum_{j=1}^{p} \left\| y^j - Ax^j \right\|^2 + \sum_{j=1}^{p} \ell\left(x^j\right)$$

- Alternating Update Algorithms (Olshausen and Field'97)
- Repeat until convergence
  - **Decoding Step**: Fix A, update X

$$x^j \leftarrow \operatorname*{argmin}_{x^j} \left\| y^j - Ax^j \right\|^2 + \ell(x^j)$$

  - **Learning Step***:* Fix X, update A

$$A \leftarrow A \ - \eta \nabla_{\mathrm{A}} \mathcal{E}(A,X)$$

Previous Works

	sparsity	Over-cmplt.	#samples
LP-based Algo [SWW'12]	$k \lesssim \sqrt{n}$	No	n^2
Combinatorial [AGM'13]][AAJNT'13]	$k \lesssim n^{1/2-\eta}$	Yes	m^2
Lassere/SoS relaxation [BKS'14]	$k \lesssim n^{1-\eta}$	Yes	$m^{poly\left(\frac{1}{\epsilon}\right)}$

Previous Works

	sparsity	Over-cmplt.	#samples
LP-based Algo [SWW'12]	$k \lesssim \sqrt{n}$	No	n^2
Combinatorial [AGM'13]][AAJNT'13]	$k \lesssim n^{1/2-\eta}$	Yes	m^2
Lassere/SoS relaxation [BKS'14]	$k \lesssim n^{1-\eta}$	Yes	$m^{poly\left(\frac{1}{\epsilon}\right)}$

These provable approaches are highly noncompetitive with non-convex optimization via alternating updates!

Previous Works

	sparsity	Over-cmplt.	#samples
LP-based Algo [SWW'12]	$k \lesssim \sqrt{n}$	No	n^2
Combinatorial [AGM'13]][AAJNT'13]	$k \lesssim n^{1/2-\eta}$	Yes	m^2
Lassere/SoS relaxation [BKS'14]	$k \lesssim n^{1-\eta}$	Yes	$m^{poly\left(\frac{1}{\epsilon}\right)}$

These provable approaches are highly noncompetitive with non-convex optimization via alternating updates!

Can we analyze non-convex optimization??

Previous Works

	sparsity	Over-cmplt.	#samples
LP-based Algo [SWW'12]	$k \lesssim \sqrt{n}$	No	n^2
Combinatorial [AGM'13]][AAJNT'13]	$k \lesssim n^{1/2-\eta}$	Yes	m^2
Lassere/SoS relaxation [BKS'14]	$k \lesssim n^{1-\eta}$	Yes	$m^{poly\left(\frac{1}{\epsilon}\right)}$

These provable approaches are highly noncompetitive with non-convex optimization via alternating updates!

Can we analyze non-convex optimization??

Local convergence is known (if start with dictionary that is $1/k$ – close to optimal) [AGM'13][AAJNT'13]

- This paper: convergence starting from $1/\log n$-close.
  + better runtime/sample complexity

Model Assumptions

$$n\text{-dim}\left\{\begin{bmatrix} y^1 & \cdots & \cdots & y^p \end{bmatrix}\right. = \begin{bmatrix} \vdots \\ A_1^* \\ \vdots \end{bmatrix} \cdots \cdots \begin{bmatrix} \vdots \\ A_m^* \\ \vdots \end{bmatrix} \cdot \begin{bmatrix} x^{*1} & \cdots & \cdots & x^{*p} \end{bmatrix}$$

data dictionary representation

Model Assumptions

$$n\text{-dim}\left\{\begin{bmatrix} y^1 & \cdots & \cdots & y^p \end{bmatrix}\right. = \begin{bmatrix} \vdots & & & \vdots \\ A_1^* & \cdots & \cdots & A_m^* \\ \vdots & & & \vdots \end{bmatrix} \cdot \begin{bmatrix} x^{*1} & \cdots & \cdots & x^{*p} \end{bmatrix}$$

data dictionary representation

- $m = O(n)$;
- $\|A_i^*\| = 1$;
- $\langle A_i^*, A_j^* \rangle \lesssim 1/\sqrt{n}$;
- $\|A^*\| \leq \sqrt{m/n}$;

Model Assumptions

$$n\text{-dim}\left\{\begin{bmatrix} y^1 & \cdots & \cdots & y^p \end{bmatrix}\right. = \begin{bmatrix} \vdots & & & \vdots \\ A_1^* & \cdots & \cdots & A_m^* \\ \vdots & & & \vdots \end{bmatrix} \cdot \begin{bmatrix} x^{*1} & \cdots & \cdots & x^{*p} \end{bmatrix}$$

data

dictionary

- $m = O(n)$;
- $\|A_i^*\| = 1$;
- $\langle A_i^*, A_j^* \rangle \lesssim 1/\sqrt{n}$;
- $\|A^*\| \leq \sqrt{m/n}$;

representation

- $k \lesssim \sqrt{n}$
- i.i.d k-sparse columns
- entries of x^{*j} has bounded correlation

Our results

Theorem 1 (Convergence):

- Given A^0 column-wise $1/\log n -$ close to A^*
- Olshausen and Field algo (with a simpler decoding) returns A^s after s iteration s.t.

$$\|A^s - A^*\|_F{}^2 \leq 0.99^s \cdot \|A^0 - A^*\|_F^2 + O(mk^2/n^2)$$

Our results

Theorem 1 (Convergence):

- Given A^0 column-wise $1/\log n -$ close to A^*
- Olshausen and Field algo (with a simpler decoding) returns A^s after s iteration s.t.

$$\|A^s - A^*\|_F{}^2 \leq 0.99^s \cdot \|A^0 - A^*\|_F^2 + O(mk^2/n^2)$$

- $\|A_i^0 - A_i^*\| \leq 1/\log n$

Our results

Theorem 1 (Convergence):

- Given A^0 column-wise $1/\log n -$ close to A^*
- Olshausen and Field algo (with a simpler decoding) returns A^s after s iteration s.t.

$$\|A^s - A^*\|_F{}^2 \leq 0.99^s \cdot \|A^0 - A^*\|_F^2 + O(mk^2/n^2)$$

- $\|A_i^0 - A_i^*\| \leq 1/\log n$

Theorem 2 (Initialization):
spectral method based algo returns A^0 that is $1/\log n -$ close to A^*

	sparsity	Over-cmplt.	#samples
LP-based Algo [SWW'12]	$k \lesssim \sqrt{n}$	No	n^2
Combinatorial [AGM'13]][AAJNT'13]	$k \lesssim n^{1/2-\eta}$	Yes	$m^2 \cdot 1/\epsilon$
Lassere/SoS relaxation [BKS'14]	$k \lesssim n^{1-\eta}$	Yes	$m^{poly\left(\frac{1}{\epsilon}\right)}$
This paper	$k \lesssim \sqrt{n}$	Yes	$\tilde{O}(mk \log 1/\epsilon)$

	sparsity	Over-cmplt.	#samples
LP-based Algo [SWW'12]	$k \lesssim \sqrt{n}$	No	n^2
Combinatorial [AGM'13]][AAJNT'13]	$k \lesssim n^{1/2-\eta}$	Yes	$m^2 \cdot 1/\epsilon$
Lassere/SoS relaxation [BKS'14]	$k \lesssim n^{1-\eta}$	Yes	$m^{poly\left(\frac{1}{\epsilon}\right)}$
This paper	$k \lesssim \sqrt{n}$	Yes	$\tilde{O}(mk \log 1/\epsilon)$

Non-convex approach is powerful!

- Aside: new initialization also heavily inspired by alt. update algorithm.

Outline

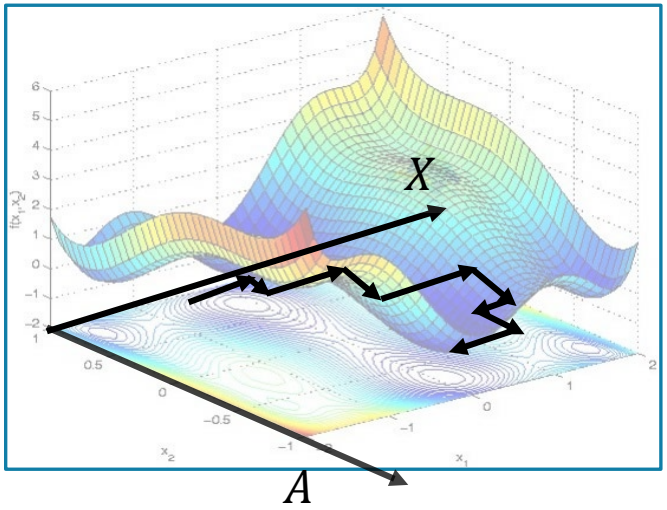
Efficient and provable algorithms for sparse coding?

- Introduction
  - Non-convex heuristics used in practice
  - Genearitve model and main results
- Analyzing alternating minimization

Alternating Updates ⟶ Updates on A

- $\min_{A,X} \mathcal{E}(A, X)$
- **Decoding Step**: Fixing A, update
- $x^j \leftarrow \underset{x^j}{\operatorname{argmin}} \left\| y^j - Ax^j \right\|^2 + \ell(x^j)$
- **Learning Step***:* Fixing X, update A

$$A \leftarrow A - \eta \nabla_A \mathcal{E}(A, X)$$

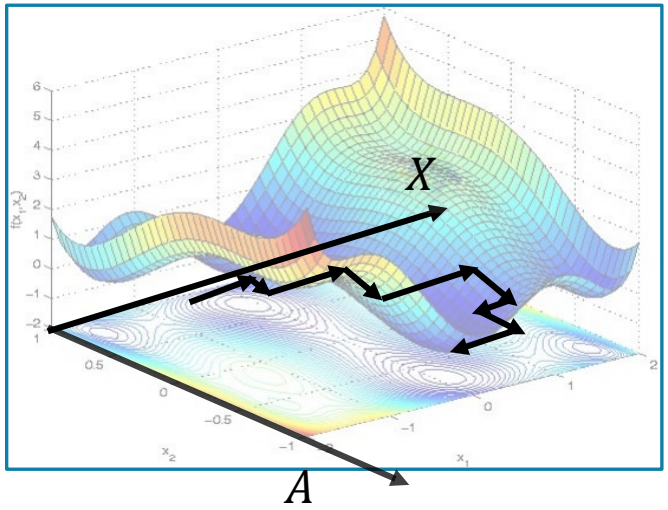


Alternating Updates ⟶ Updates on A

- $\min_{A,X} \mathcal{E}(A, X)$
- **Decoding Step**: Fixing A, update
- $x^j \leftarrow \underset{x^j}{\operatorname{argmin}} \left\| y^j - Ax^j \right\|^2 + \ell(x^j)$
- **Learning Step**: Fixing X, update A

$$A \leftarrow A - \eta \nabla_A \mathcal{E}(A, X)$$

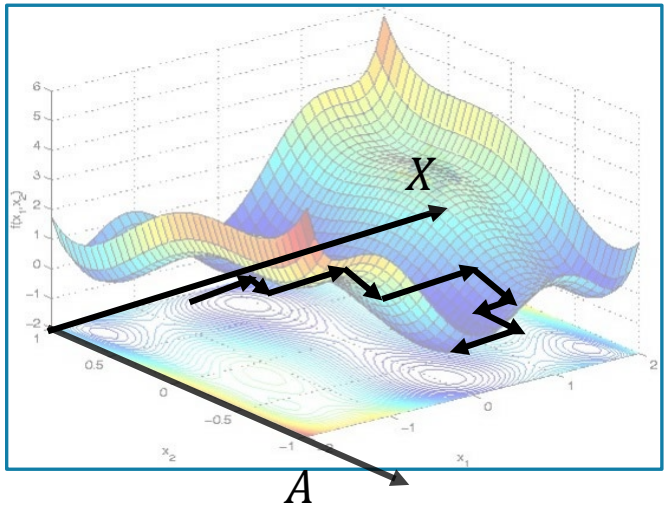
$$A^{s+1} = A^s - \eta g^s$$



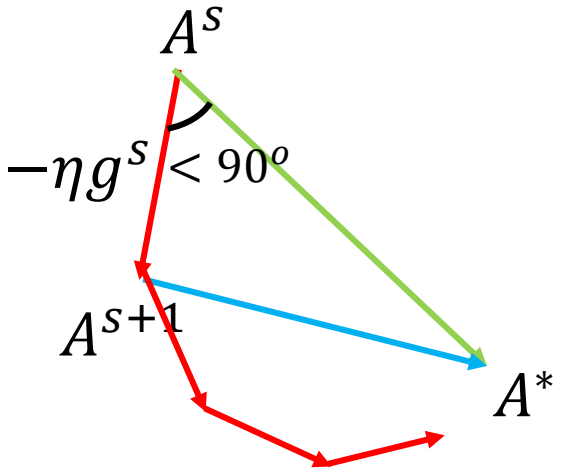
Alternating Updates ⟶ Updates on A

- $\min_{A,X} \mathcal{E}(A, X)$
- **Decoding Step**: Fixing A, update
- $x^j \leftarrow \underset{x^j}{\operatorname{argmin}} \left\| y^j - Ax^j \right\|^2 + \ell(x^j)$
- **Learning Step**: Fixing X, update A

$$A \leftarrow A - \eta \nabla_A \mathcal{E}(A, X)$$



$$A^{s+1} = A^s - \eta g^s$$



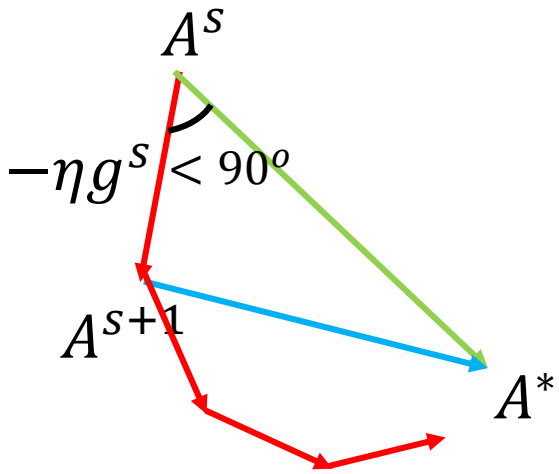
Proof Framework

- Suppose we have an update rule

$$A^{s+1} = A^s - \eta g^s$$

and a desired point A^*

(g doesn't have to be grad; and there may be no "objective fn")



Proof Framework

- Suppose we have an update rule

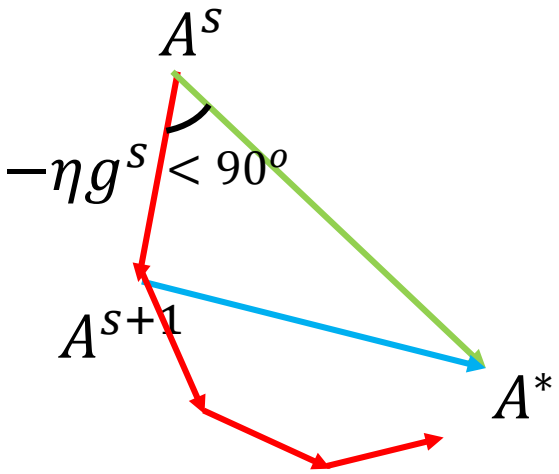
$$A^{s+1} = A^s - \eta g^s$$

and a desired point A^*

(g doesn't have to be grad; and there may be no "objective fn")

Condition:

$$\langle g^s, A^s - A^* \rangle \geq \alpha \|A^s - A^*\|_F^2 + \beta \|g^s\|_F^2 - \epsilon_s$$



Proof Framework

- Suppose we have an update rule

$$A^{s+1} = A^s - \eta g^s$$

and a desired point A^*

(g doesn't have to be grad; and there may be no "objective fn")

Condition:

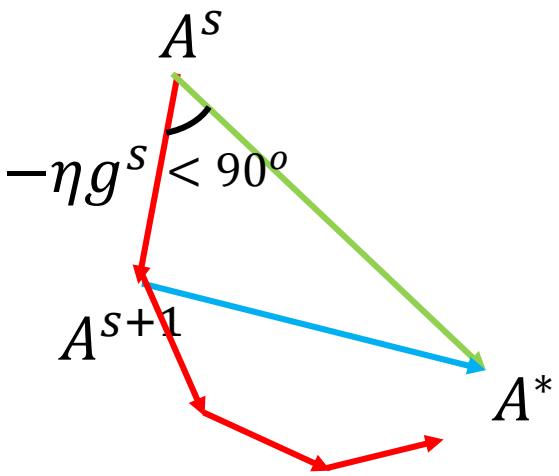
$$\langle g^s, A^s - A^* \rangle \geq \alpha \|A^s - A^*\|_F^2 + \beta \|g^s\|_F^2 - \epsilon_s$$

$$\Downarrow$$

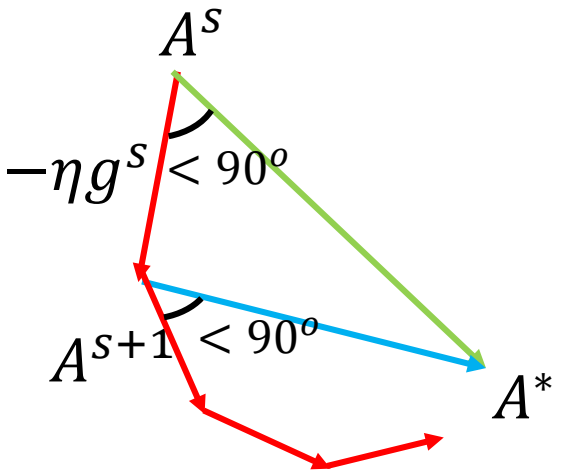
geometrical convergence:

$$\|A^s - A^*\|_F^2 \leq (1 - 2\alpha\eta)^s \cdot \|A^0 - A^*\|_F^2 + \epsilon/\alpha$$

(where $\eta \leq 2\beta, \epsilon = \max \epsilon_s$)

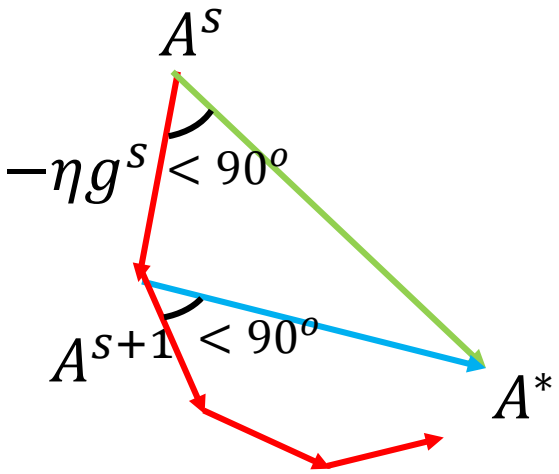


Analysis of Sparse Coding



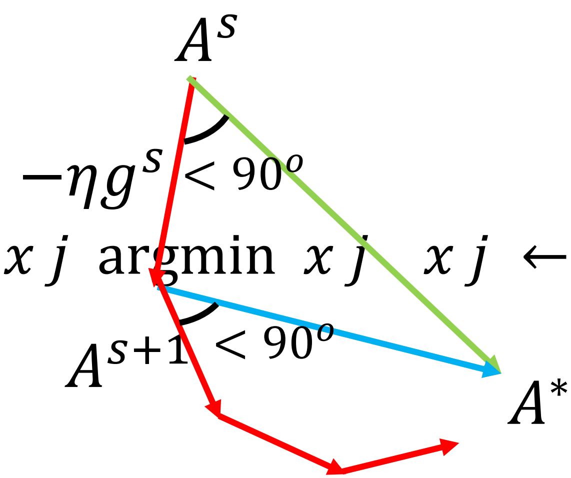
Analysis of Sparse Coding

- $\min_{A,X} \mathcal{E}(A, X)$



Analysis of Sparse Coding

- argmin $x\,j$ $x\,j$ $x\,j$ $x\,j$ argmin $x\,j$ $x\,j \leftarrow$
 argmin $x\,j$ $y\,j - A\,x\,j$ 2 $+$
 $\ell(x\,j)$ $y\,j - A\,x\,j$ 2 $y\,j$ $-$
 $A\,x\,j$ $y\,j$ $y\,j$ $y\,j -$
 $jjyyjjxxnimgr$ argmin $x\,j$ a argmin $x\,j$ $x\,j$ $x\,j$ $x\,j$ argmin $x\,j$ $x\,j \leftarrow$
 argmin $x\,j$ $y\,j - A\,x\,j$ 2 $+$
 $\ell(x\,j)$ $y\,j - A\,x\,j$ 2 $y\,j$ $-$
 $A\,x\,j$ $y\,j$ $y\,j$ $y\,j -$
 $jjyyjjxxnimgr$ argmin $x\,j$ a argmin $x\,j$ $x\,j$ $x\,j$ $x\,j$ argmin $x\,j$ $x\,j \leftarrow$
 argmin $x\,j$ $y\,j - A\,x\,j$ 2 $+$
 $\ell(x\,j)$ $y\,j - A\,x\,j$ 2 $y\,j$ $-$
 $A\,x\,j$ $y\,j$ $y\,j$ $y\,j -$
 $jjyyjjxxnimgr$ argmin $x\,j$ a argmin $x\,j$ $x\,j$ $x\,j$ $x\,j$ argmin $x\,j$ $x\,j \leftarrow$
 argmin $x\,j$ $y\,j - A\,x\,j$ 2 $+$
 $\ell(x\,j)$ $y\,j - A\,x\,j$ 2 $y\,j$ $-$
 $A\,x\,j$ $y\,j$ $y\,j$ $y\,j -$
 $jjyyjjxxnimgr$ argmin $x\,j$ $aXX\,A,X$ min A,X $\mathcal{E}\,A,X$

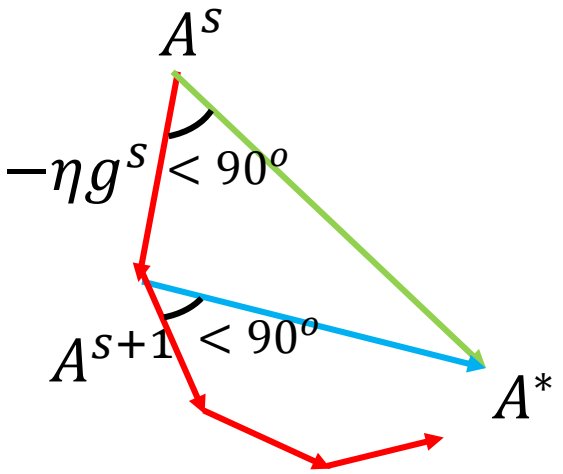


- **Decoding Step**: Fixing A, update

Analysis of Sparse Coding

- $\min_{A,X} \mathcal{E}(A, X)$
- **Decoding Step**: Fixing A, update
- $x^j \leftarrow$ $x^j \leftarrow \text{threshold}(A^T y^j)$
- **Learning Step**: Fixing X, update A

$$A \leftarrow A - \eta g$$



Proof sketch for convergence Thm:

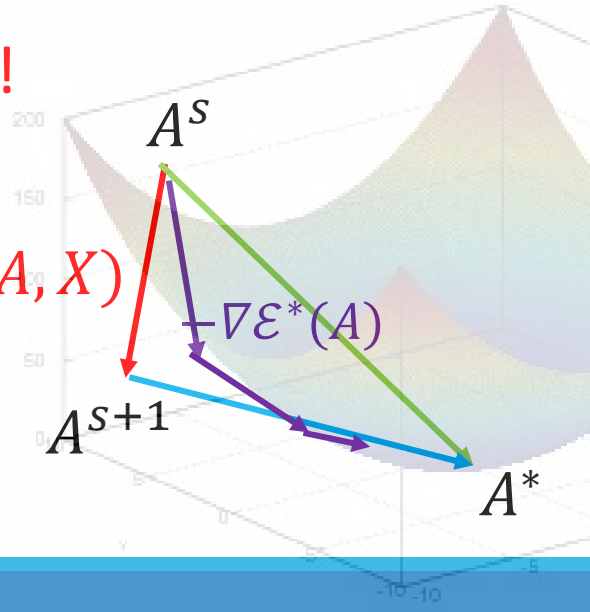
- Get a "closed" form for $g \approx B_A A - C_A A^* + \epsilon$ with $B_A, C_A \approx I$
  - using the randomness of data y^j
  - using incoherence and isotropy of A^*
  - + other tricks
- Check that $\langle g, A - A^* \rangle \geq \alpha \|A - A^*\|^2 + \beta \|g\|^2 - \epsilon_s$

Why should we expect that $g = \nabla_A \mathcal{E}(A, X)$ is correlated with $A - A^*$?

- Consider unknown convex function

$$\mathcal{E}^*(A) \coloneqq \mathcal{E}(A, X^*) = \sum_{j=1}^{p} \left\| y^j - A x^{*j} \right\|^2 + C$$

- $X \approx X^* \Longrightarrow \nabla_A \mathcal{E}(A, X) \approx \nabla_A \mathcal{E}(A, X^*) = \nabla \mathcal{E}^*(A)$
- $\mathcal{E}^*(\cdot)$ is convex $\Longrightarrow \nabla \mathcal{E}^*(A)$ correlated with $A - A^*$
- Alt. update is simulating grad descent on $\mathcal{E}^*$!
- (Actual proof doesn't use this explicitly)



Discussion and Open Questions

- Generalizable to alternating updates algo. for other hidden variable models?
  - Key technical difficulty: what if the decoding doesn't have a simple closed form?

- Limited to "local" convergence analysis
  - Analyzing global convergence from random initialization?

- Practical algo. (initialization) beyond $k = \sqrt{n}$?