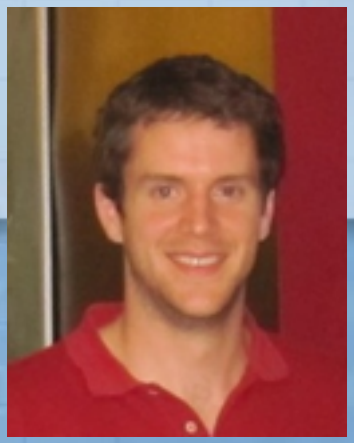
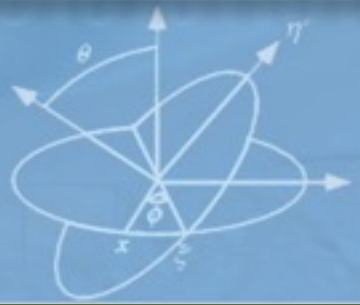
JHU vision lab

Global Optimality in Matrix and Tensor Factorizations, Deep Learning and More



Ben Haeffele and René Vidal
Center for Imaging Science
Institute for Computational Medicine

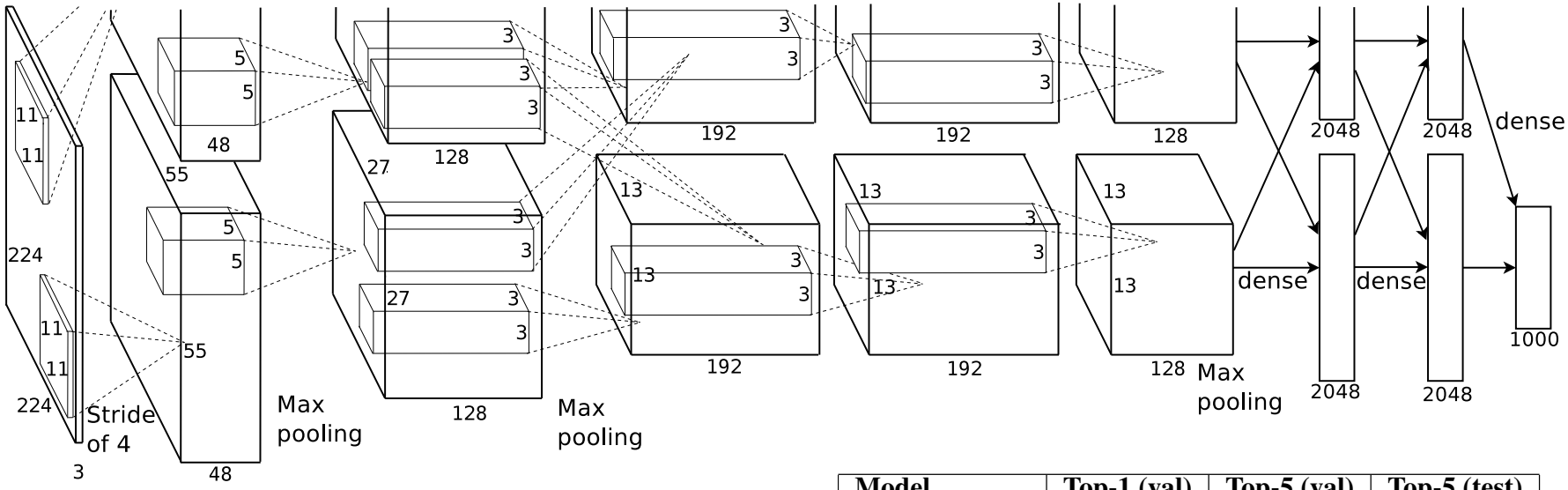


THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins

Center for IMAGING SCIENCE

Learning Deep Image Feature Hierarchies

- Deep learning gives ~ 10% improvement on ImageNet
  - 1.2 million images, 1000 categories, 60 million parameters



Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

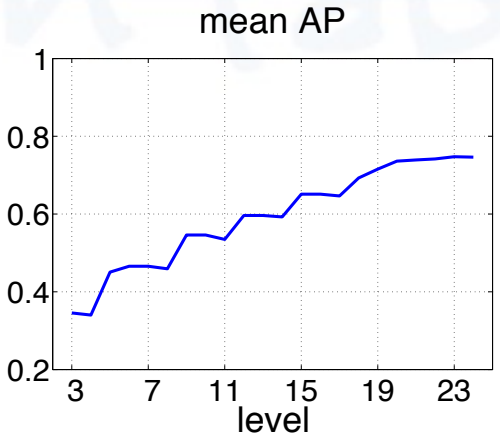
Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were "pre-trained" to classify the entire ImageNet 2011 Fall release. See Section 6 for details.

[1] Krizhevsky, Sutskever and Hinton. ImageNet classification with deep convolutional neural networks, NIPS'12.
[2] Sermanet, Eigen, Zhang, Mathieu, Fergus, LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. ICLR'14.
[3] Donahue, Jia, Vinyals, Hoffman, Zhang, Tzeng, Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. ICML'14.

Why These Improvements in Performance?

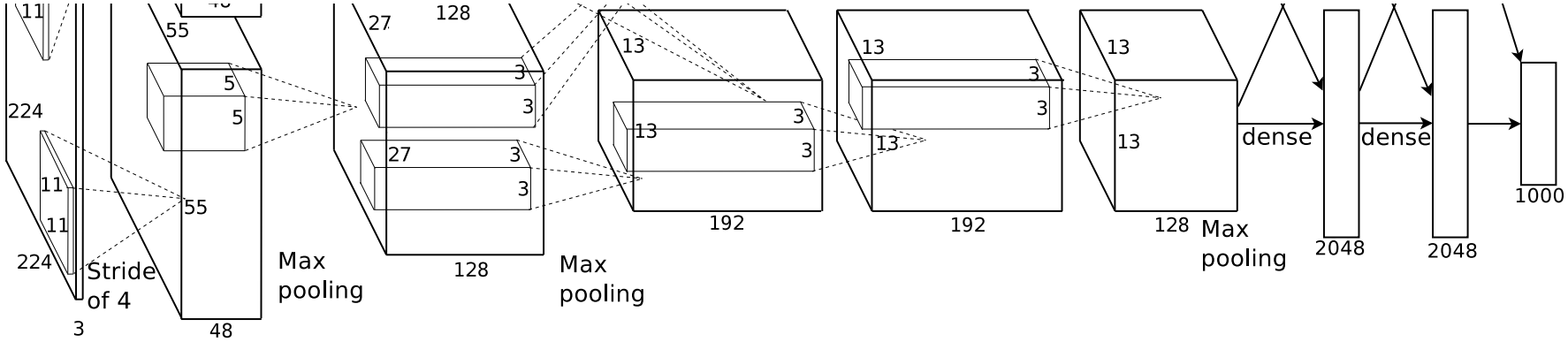
- More layers [1]
  - Multiple layers capture more invariances
  - Features are learned rather than hand-crafted
- More data
  - There is more data to train deeper networks
- More computing
  - GPUs go hand in hand with learning methods
- First attempt at a theoretical justification of invariance [2,3]
  - Theoretical support for invariance via scattering transform
  - Each layer must be a contraction to keep data volume bounded
  - Optimization issues are not discussed: stage-wise learning is used



[1] Razavian, Azizpour, Sullivan, Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition. CVPRW'14.
[2] J Bruna, S Mallat, Invariant scattering convolution networks.
[3] J Bruna, S Mallat. Classification with scattering operators, arXiv preprint arXiv:1011.3023, 2010
[4] Mallat and Waldspurger. Deep Learning by Scattering, arXiv 2013

What About Optimization?

- The learning problem is non-convex



$$\Phi(X^1, \ldots, X^K) = \psi_K(\cdots \psi_2(\psi_1(VX^1)X^2)\cdots X^K)$$

nonlinearity features weights

$$\min_{X^1,\ldots,X^K} \ell(Y, \Phi(X^1, \ldots, X^K)) + \lambda\Theta(X^1, \ldots, X^K)$$

loss labels regularizer

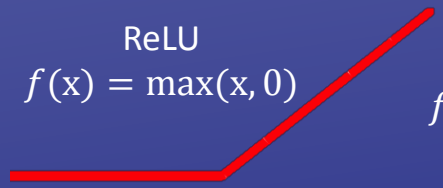
What About Optimization?

THE NON-LINEAR PART

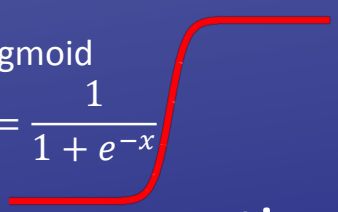
- Usually $\psi = g \circ f$.

- f is the (point-wise) activation function

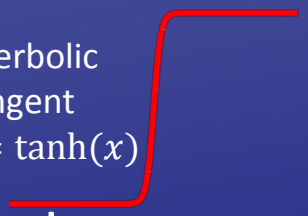
ReLU
$f(\mathrm{x}) = \max(\mathrm{x}, 0)$



Sigmoid
$f(x) = \frac{1}{1+e^{-x}}$



Hyperbolic tangent
$f(x) = \tanh(x)$



- g is a pooling or an aggregation operator.

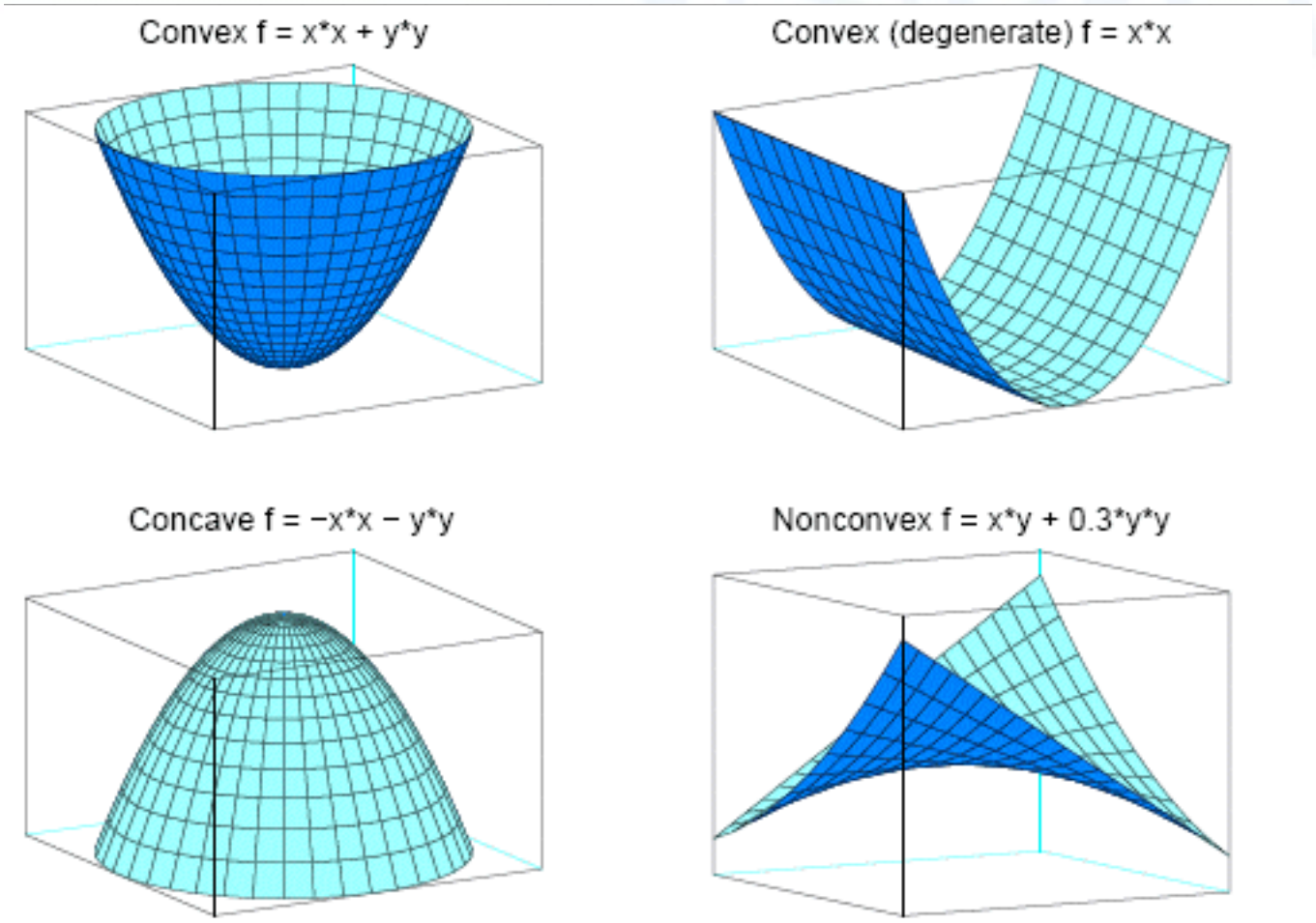
Max pooling
$\max_i V_i$

Mean pooling
$\frac{1}{n}\sum_{i=1}^{n} V_i$

l_p pooling
$\sqrt[p]{\sum_{i=1}^{n} V_i^p}$

Credit to Raja Giryes and Guillermo Sapiro

Why do We Care About Convexity?



- A local minimizer of a convex problem is a global minimizer.

http://support.sas.com/documentation/cdl/en/ormpug/63352/HTML/default/viewer.htm#ormpug_optqp_sect001.htm

Why is Non Convexity a Problem?

How is Non Convexity Handled?

- The learning problem is non-convex

$$\min_{X^1,\ldots,X^K} \ell(Y, \Phi(X^1,\ldots,X^K)) + \lambda\Theta(X^1,\ldots,X^K)$$

  - Back-propagation, alternating minimization, descent method

- To get a good local minima
  - Random initialization
  - If training error does not decrease fast enough, start again
  - Repeat multiple times

- Mysteries
  - One can find many solutions with similar objective values
  - Rectified linear units work better than sigmoid/hyperbolic tangent
  - Dead units (zero weights)

Related Work

- **Gradient descent on square loss**
  - No spurious local optima for linear networks: Baldi & Hornik '89
  - Failure cases: manifold of spurious local optima. Frasconi '97
  - Random first layer weights suffice for polynomials under gaussian input: Andoni et al. '14
  - Incremental training with polynomial activations. Livni et al. '14

- **Models with stochastic weights and inputs**
  - Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non- convex optimization," NIPS, 2014.
  - A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun, "The Loss Surfaces of Multilayer Networks," AISTAT, 2015.
- **Stochastic models trained using moment methods**
  - M Janzamin, H Sedghi, A Janzamin, Beating the Perils of Non-Convexity: Guaranteed Training of Neural Networks using Tensor Methods

P Baldi, K Hornik, Neural networks and principal component analysis: Learning from examples without local minima, Neural networks, 1989.
M Brady, R Raghavan, J Slawny. Back propagation fails to separate where perceptrons succeed. IEEE Trans Circuits & Systems, 36(5):665–674, 1989.
M Gori, A Tesi. On the problem of local minima in backpropagation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 14(1):76–86, 1992.
P Frasconi, M Gori, and A Tesi. Successes and failures of backpropagation: A theoretical. Progress in Neural Networks: Architecture, 5:205, 1997.

Contributions

$$\min_{X^1,\ldots,X^K} \ell(Y, \Phi(X^1,\ldots,X^K)) + \lambda\Theta(X^1,\ldots,X^K)$$

- **Assumptions:**
  - $\ell(Y, X)$: convex and once differentiable in X
  - Φ and Θ: sums of positively homogeneous functions of same degree

$$f(\alpha X^1,\ldots,\alpha X^K) = \alpha^p f(X^1,\ldots,X^K) \quad \forall \alpha \geq 0$$

- **Theorem 1:** A local minimizer such that for some *i* and all k $X_i^k = 0$ is a global minimizer

- **Theorem 2:** If the size of the network is large enough, local descent can reach a global minimizer from any initialization

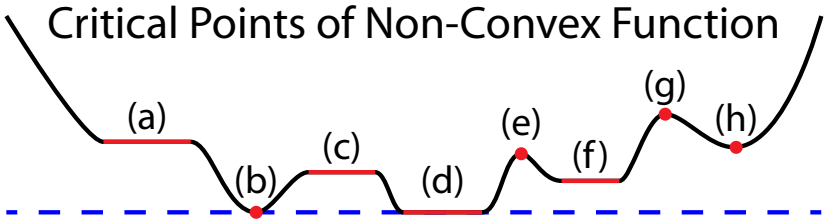
Contributions

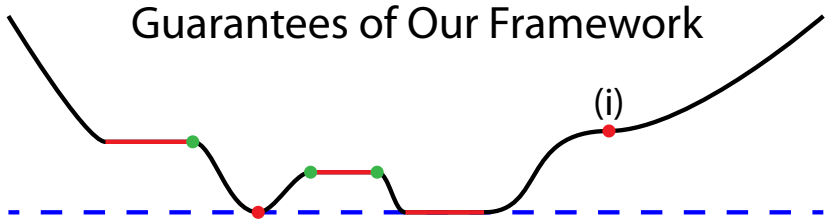
$$\min_{X^1,\ldots,X^K} \ell(Y, \Phi(X^1,\ldots,X^K)) + \lambda\Theta(X^1,\ldots,X^K)$$

- **Assumptions:**
  - $\ell(Y, X)$: convex and once differentiable in X
  - Φ and Θ: sums of positively homogeneous functions of same degree

$$f(\alpha X^1,\ldots,\alpha X^K) = \alpha^p f(X^1,\ldots,X^K) \quad \forall \alpha \geq 0$$

- **Theorem 2:**





Outline

- Global Optimality in Structured Matrix Factorization [1,2]
  - PCA, Robust PCA, Matrix Completion
  - Nonnegative Matrix Factorization
  - Dictionary Learning
  - Structured Matrix Factorization

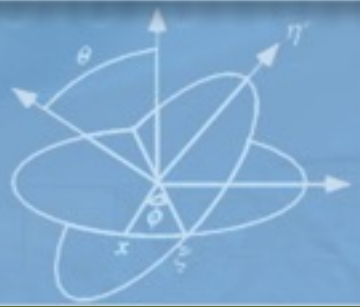
- Global Optimality in Positively Homogeneous Factorization [2]
  - Tensor Factorization
  - Deep Learning
  - More

[1] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14
[2] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv, '15

JHU vision lab

Global Optimality in Structured Matrix Factorization

René Vidal
Center for Imaging Science
Institute for Computational Medicine

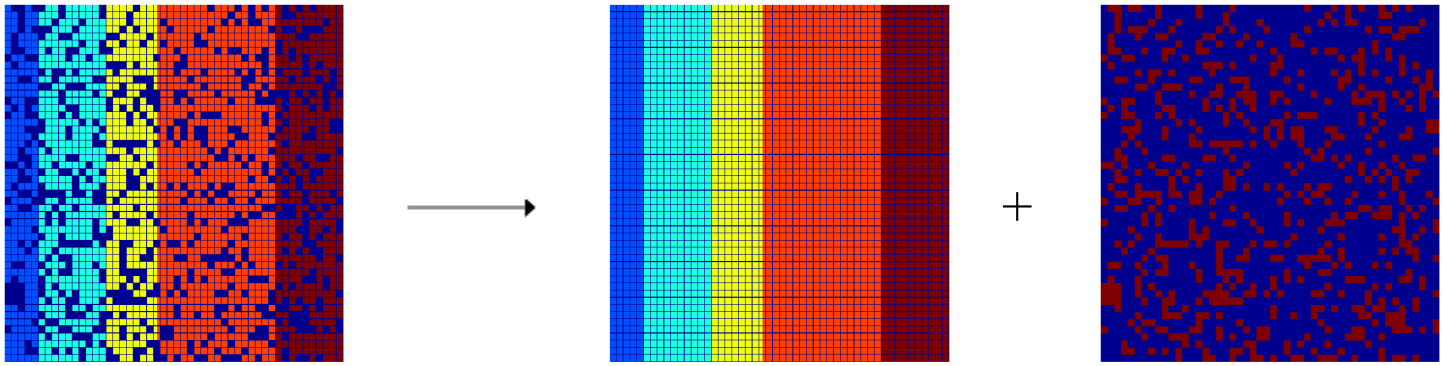


Low Rank Modeling

- Models involving factorization are ubiquitous
  - PCA
  - Nonnegative Matrix Factorization
  - Dictionary Learning
  - Matrix Completion
  - Robust PCA

$$\|X\|_* = \sum \sigma_i(X)$$

$$\min_X \ \|Y - X\|_1 + \lambda \|X\|_*$$

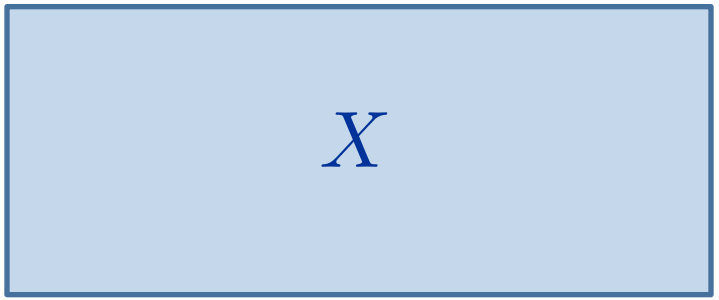


http://perception.csl.illinois.edu/matrix-rank/home.html

Typical Low-Rank Formulations

- Convex formulations

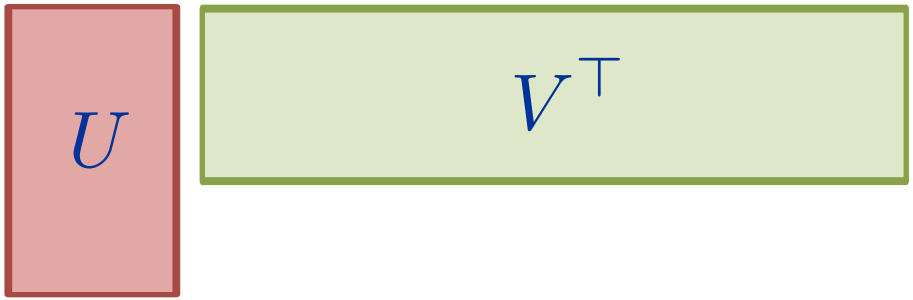
$$\min_X \ell(Y, X) + \lambda\Theta(X)$$



  - Robust PCA
  - Matrix Completion

- Factorized formulations

$$\min_{U,V} \ell(Y, UV^\top) + \lambda\Theta(U, V)$$



  - Nonnegative matrix factorization
  - Dictionary learning

- Convex
- Large problem size
- Unstructured factors

- Non-Convex
- Small problem size
- Structured factors

Why Do We Need Structured Factors?

- Given a low-rank video $Y \in \mathbb{R}^{p \times t}$ $\min_X \|Y - X\|_1 + \lambda \|X\|_*$

(a) Original frames

(b) Low-rank $\hat{L}$

(c) Sparse $\hat{S}$

$$\min_{U,V} \ell(Y, UV^\top) + \lambda \Theta(U, V)$$

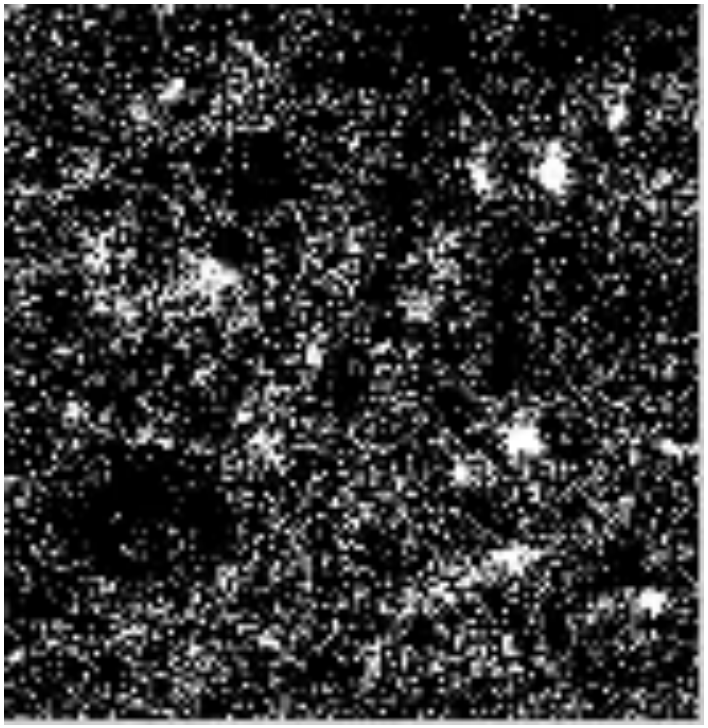
- U: spatial basis
  - Low total-variation
  - Non-negative
- V: temporal basis
  - Sparse on particular basis set
  - Non-negative

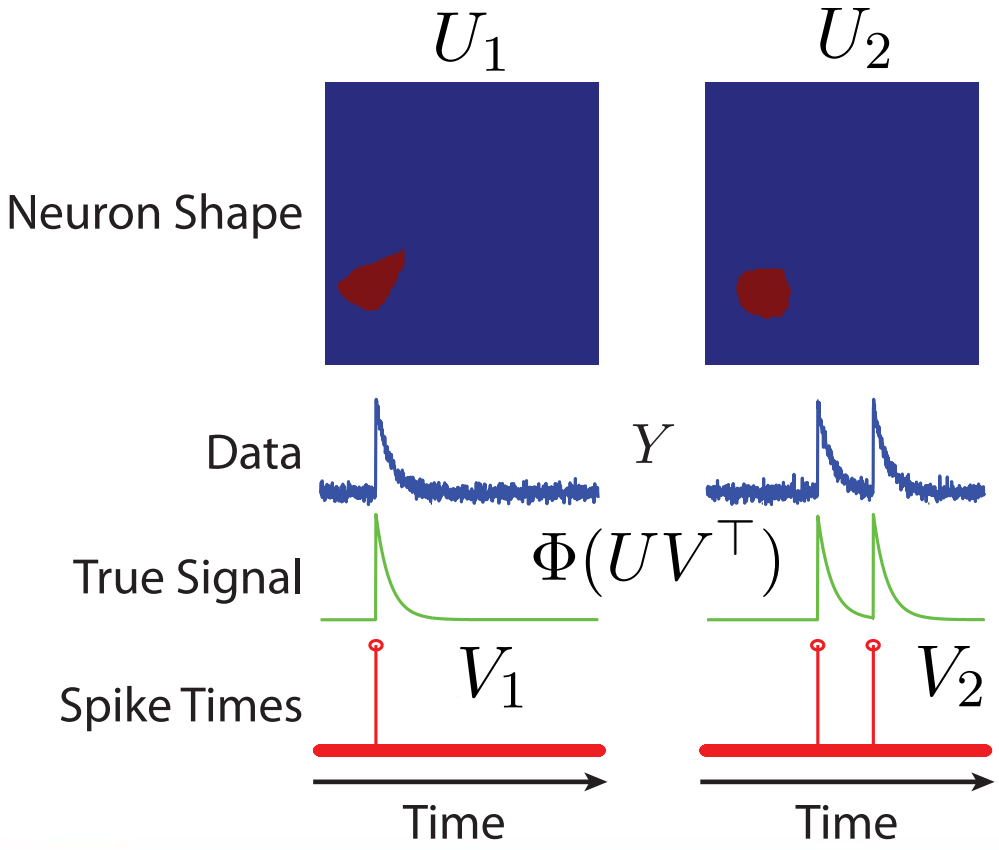
[1] Candes, Li, Ma, Wright. Robust Principal Component Analysis? Journal of the ACM, 2011.

Why Do We Need Structured Factors?

- Find neuronal shapes and spike trains in calcium imaging

$$\min_{U,V} \ell(Y, \Phi(U,V)) + \lambda\Theta(U,V)$$





Why Do We Need Structured Factors?

- Nonnegative matrix factorization

$$\min_{U,V} \|Y - UV^\top\|_F^2 \quad \text{s.t.} \quad U \geq 0, V \geq 0$$

- Sparse dictionary learning

$$\min_{U,V} \|Y - UV^\top\|_F^2 \quad \text{s.t.} \quad \|U_i\|_2 \leq 1, \|V_i\|_0 \leq r$$

- **Challenges to state-of-the-art methods**
  - Need to pick size of U and V a priori
  - Alternate between U and V, without guarantees of convergence to a global minimum

Tackling Non-Convexity: Nuclear Norm Case

- Convex problem

$$\min_X \ell(Y, X) + \lambda \|X\|_*$$

Factorized problem

$$\min_{U,V} \ell(Y, UV^\top) + \lambda \Theta(U, V)$$

- Variational form of the nuclear norm

$$\|X\|_* = \min_{U,V} \sum_{i=1}^{r} |U_i|_2 |V_i|_2 \quad \text{s.t.} \quad UV^\top = X$$

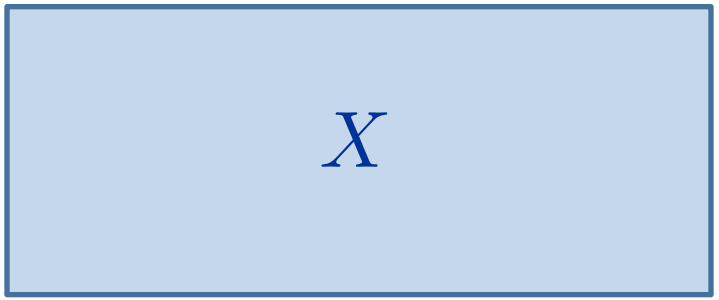
- **Theorem:** Assume loss ℓ is convex and once differentiable in X. A local minimizer of the factorized problem such that for some i $U_i = V_i = 0$ is a global minimizer of both problems

- **Intuition:** regularizer Θ "comes from a convex function"

Tackling Non-Convexity: Nuclear Norm Case

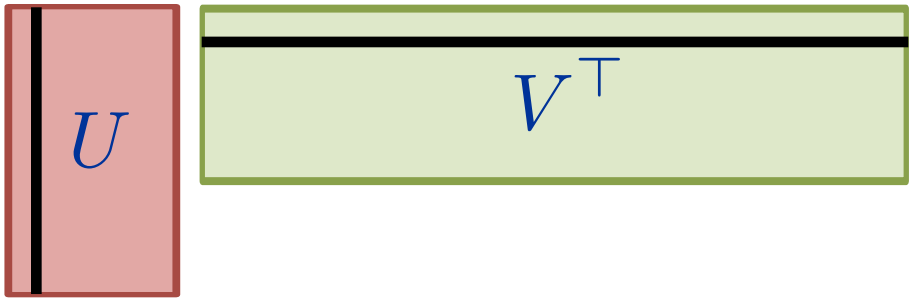
- Convex problem

$$\min_X \ell(Y, X) + \lambda \|X\|_*$$



Factorized problem

$$\min_{U,V} \ell(Y, UV^\top) + \lambda \Theta(U, V)$$



- **Theorem:** Assume loss ℓ is convex and once differentiable in X. A local minimizer of the factorized problem such that for some i $U_i = V_i = 0$ is a global minimizer of both problems

Tackling Non-Convexity: Tensor Norm Case

- A natural generalization is the projective tensor norm [1,2]

$$\|X\|_{u,v} = \min_{U,V} \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v \quad \text{s.t.} \quad UV^\top = X$$

- **Theorem 1 [3,4]:** A local minimizer of the factorized problem

$$\min_{U,V} \ell(Y, UV^\top) + \lambda \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v$$

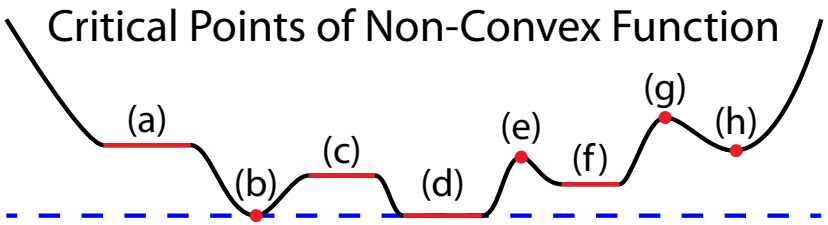
such that for some i $U_i = V_i = 0$, is a global minimizer of both the factorized problem and of the convex problem

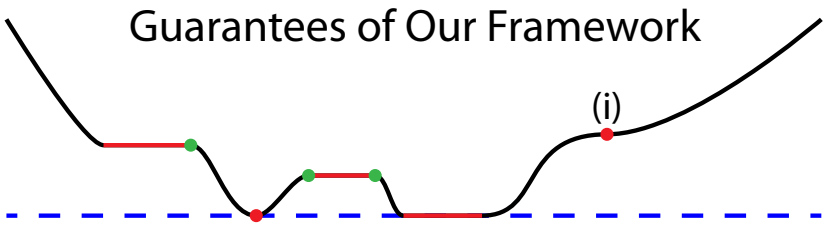
$$\min_{X} \ell(Y, X) + \lambda \|X\|_{u,v}$$

[1] Bach, Mairal, Ponce, Convex sparse matrix factorizations, arXiv 2008.
[2] Bach. Convex relaxations of structured matrix factorizations, arXiv 2013.
[3] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14
[4] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv '15

Tackling Non-Convexity: Tensor Norm Case

- **Theorem 2:** If the number of columns is large enough, local descent can reach a global minimizer from any initialization





- **Meta-Algorithm:**
  - If not at a local minima, perform local descent to reach a local minima
  - If optimality condition is satisfied, then local minima is global
  - If condition fails, choose descent direction (u,v), and set

$$r \leftarrow r + 1 \quad U \leftarrow \begin{bmatrix} U & u \end{bmatrix} \quad V \leftarrow \begin{bmatrix} V & v \end{bmatrix}$$

[1] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv '15

Example: Nonnegative Matrix Factorization

- Original formulation

$$\min_{U,V} \|Y - UV^\top\|_F^2 \quad \text{s.t.} \quad U \geq 0, V \geq 0$$

- New factorized formulation

$$\min_{U,V} \|Y - UV^\top\|_F^2 + \lambda \sum_i |U_i|_2 |V_i|_2 \quad \text{s.t.} \quad U, V \geq 0$$

  - Note: regularization limits the number of columns in (U,V)

Example: Sparse Dictionary Learning

- Original formulation

$$\min_{U,V} \|Y - UV^\top\|_F^2 \quad \text{s.t.} \quad \|U_i\|_2 \leq 1, \|V_i\|_0 \leq r$$

- New factorized formulation

$$\min_{U,V} \|Y - UV^\top\|_F^2 + \lambda \sum_i |U_i|_2 (|V_i|_2 + \gamma |V_i|_1)$$

Non Example: Robust PCA

- Original formulation [1]

$$\min_{X,E} \|E\|_1 + \lambda \|X\|_* \quad \text{s.t.} \quad Y = X + E$$

- Equivalent formulation

$$\min_{X} \|Y - X\|_1 + \lambda \|X\|_*$$

- New factorized formulation

$$\min_{U,V} \|Y - UV^\top\|_1 + \lambda \sum_i |U_i|_2 |V_i|_2$$

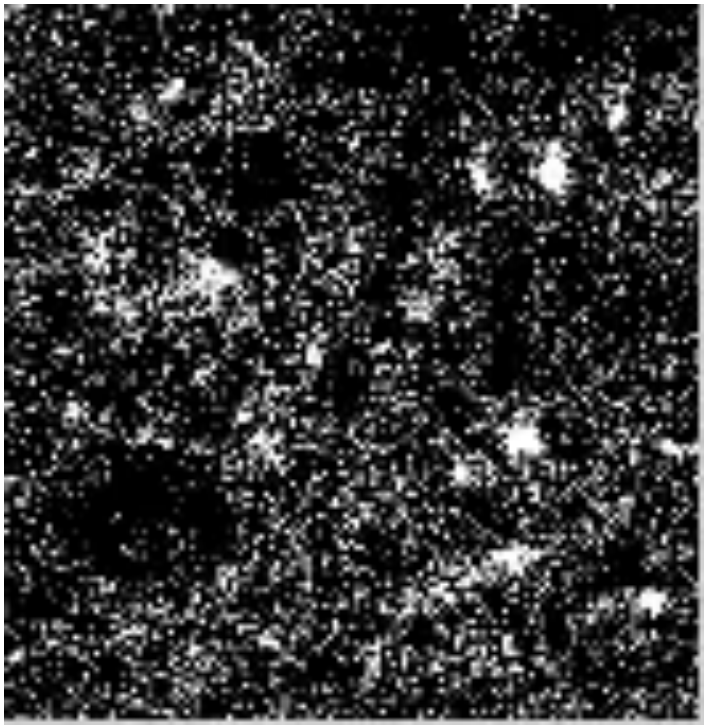
- Not an example because loss is not differentiable

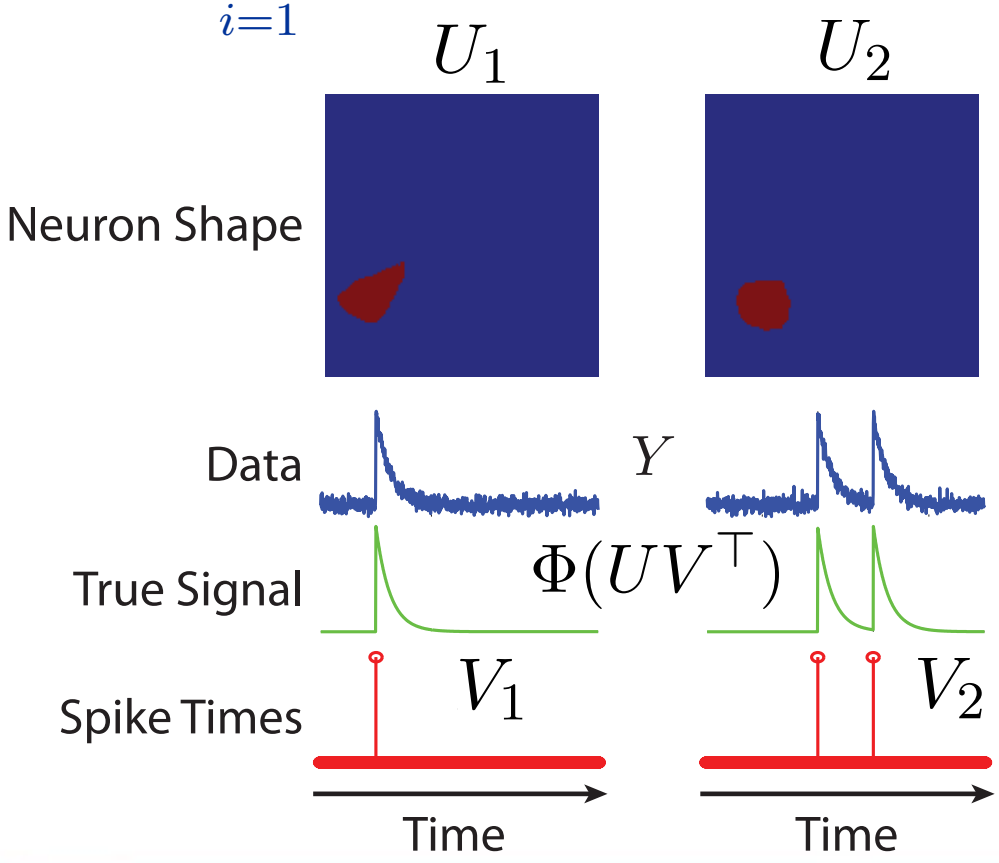
[1] Candes, Li, Ma, Wright. Robust Principal Component Analysis? Journal of the ACM, 2011.

Neural Calcium Image Segmentation

- Find neuronal shapes and spike trains in calcium imaging

$$\min_{U,V} \|Y - \Phi(UV^\top)\|_F^2 + \lambda \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v$$





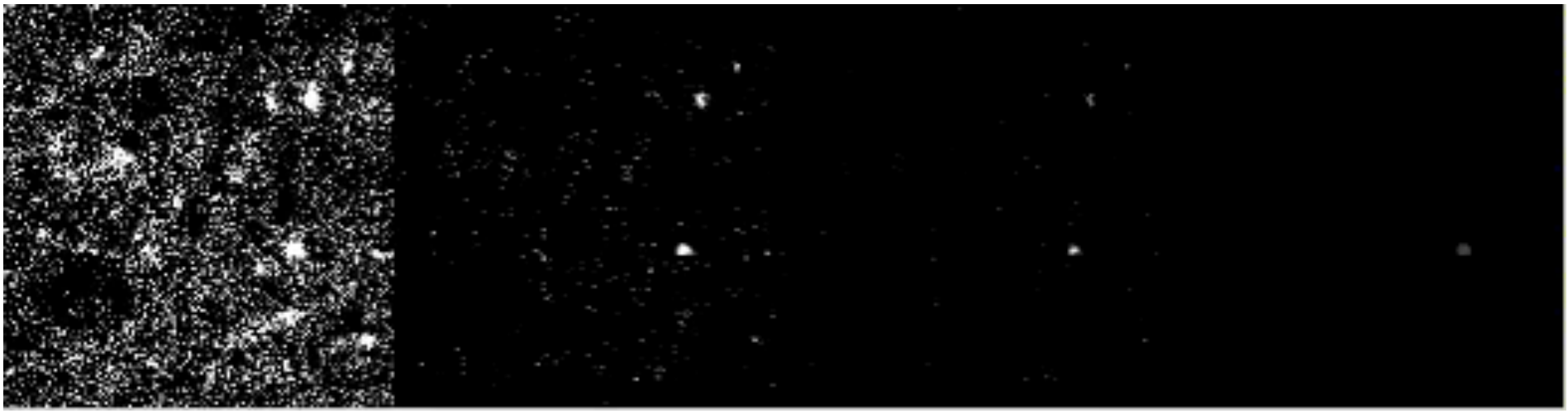
In Vivo Results (Small Area)

$$\min_{U,V} \|Y - \Phi(UV^\top)\|_F^2 + \lambda \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v$$

$$\|\cdot\|_u = \|\cdot\|_2 + \|\cdot\|_1 + \|\cdot\|_{TV}$$

$$\|\cdot\|_v = \|\cdot\|_2 + \|\cdot\|_1$$

60 microns

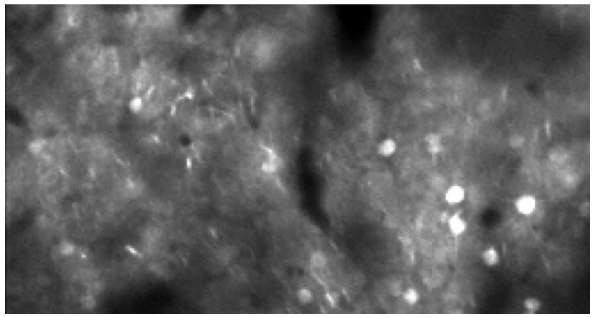


Raw Data | Sparse | + Low Rank | +Total Variation

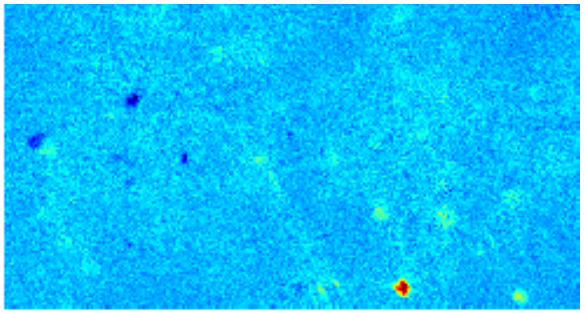
In Vivo Results

- PCA
  - Sensitive to noise
  - Hard to interpret

Mean Fluorescence

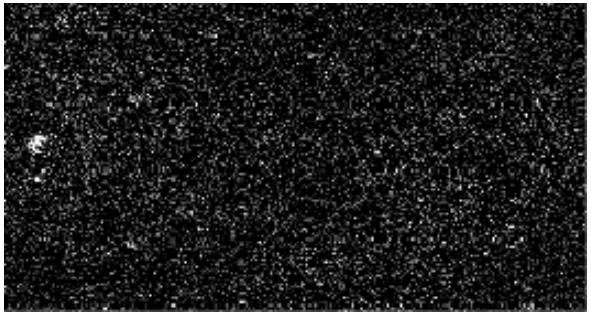


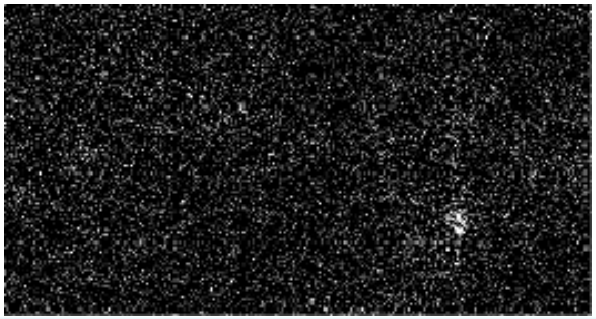
Feature obtained by PCA



- Proposed method
  - Found 46/48 manually identified active regions
  - Features are easy to interpret
  - Minimal post-processing for segmentation

Example Image Frames

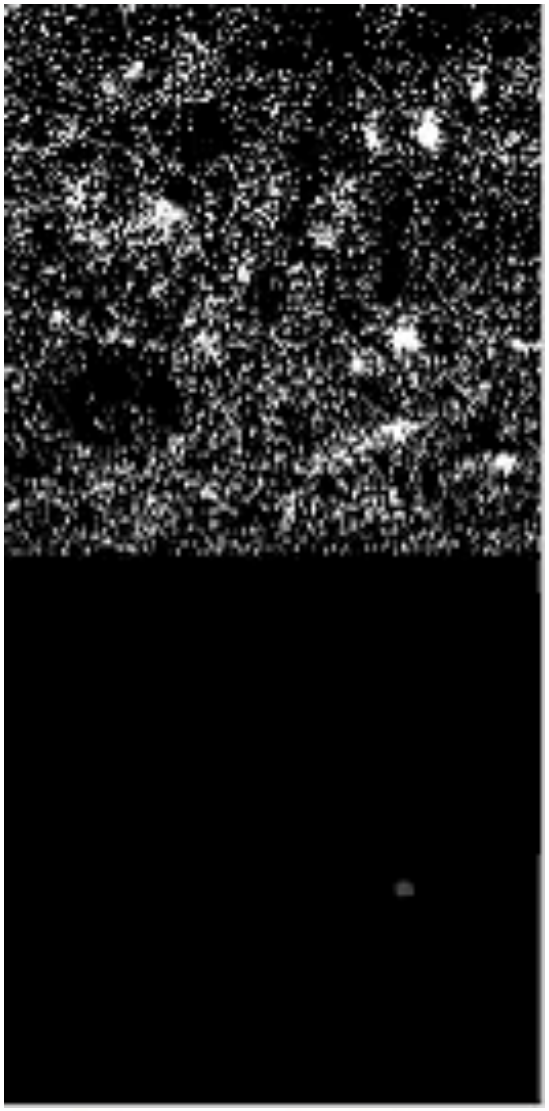


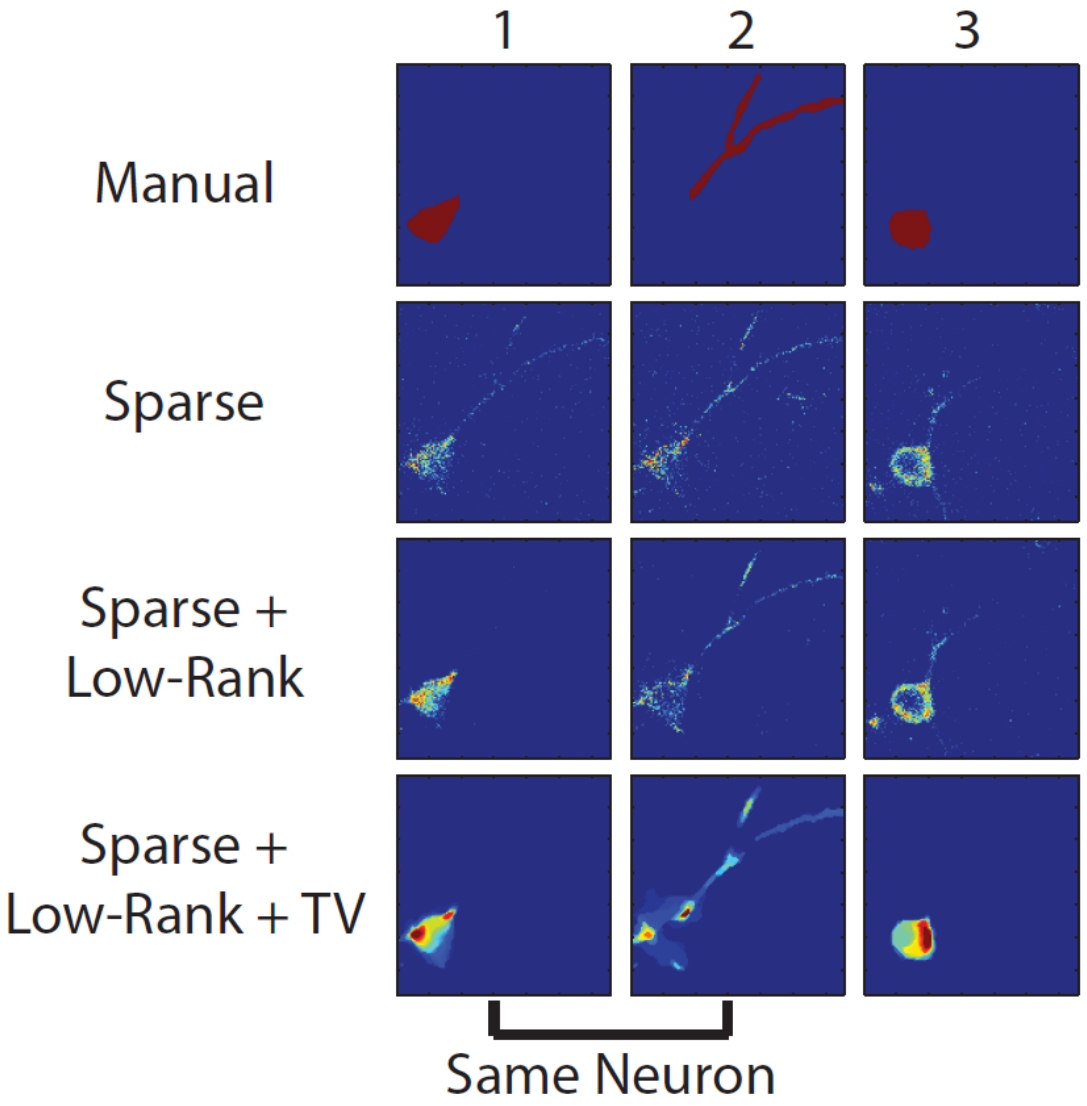


Features by Our Method

In Vivo Results (Large Area)

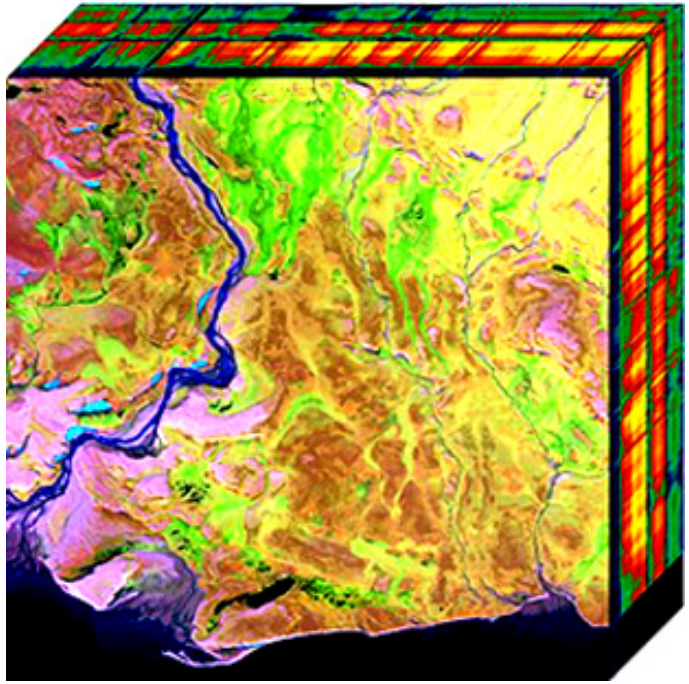
Neural Calcium Image Segmentation





Hyperspectral Compressed Recovery

- $Y \in \mathbb{R}^{p \times t}$: hyperspectral image of a certain area at multiple (t>100) wavelengths of light



- Different regions in space correspond to different materials
  - rank(Y) = number of materials
- U: spatial features
  - Low total-variation
  - Non-negative
- V: spectral features
  - Non-negative

$$\min_{U,V} \ell(Y, UV^\top) + \lambda \Theta(U, V)$$

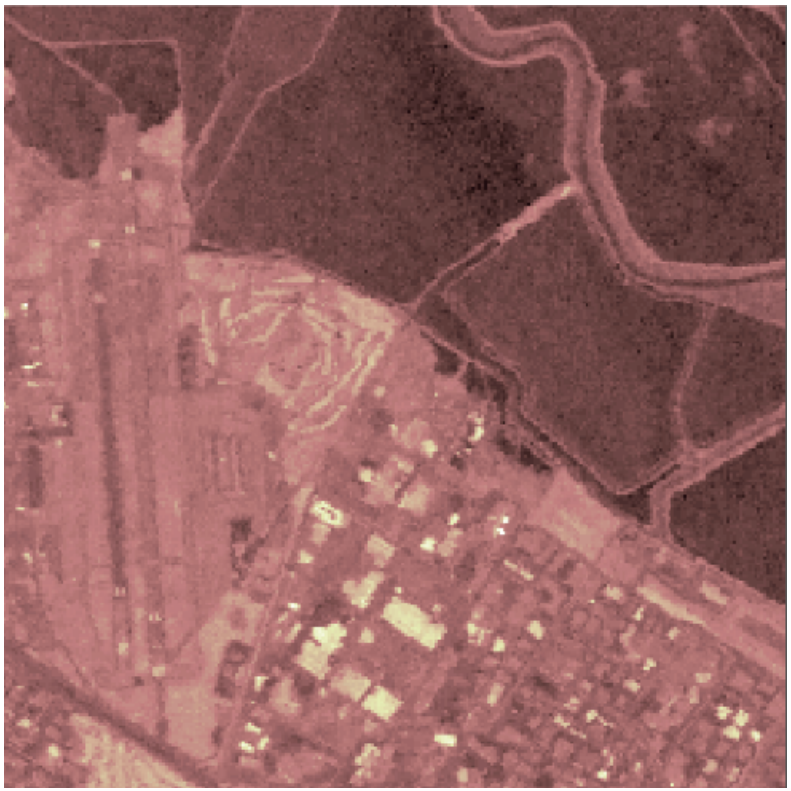
[1] Candes, Li, Ma, Wright. Robust Principal Component Analysis? Journal of the ACM, 2011.

Hyperspectral Compressed Recovery

- Prior method: NucTV (Golbabaee et al., 2012)

$$\min_X \|X\|_* + \lambda \sum_{i=1}^{t} \|X_i\|_{TV} \quad \text{s.t.} \quad \|Y - \Phi(X)\|_F^2 \leq \epsilon$$

- 180 Wavelengths
- 256 x 256 Images
- Computation per Iteration
  - SVT of whole image volume
  - 180 TV Proximal Operators
  - Projection onto Constraint Set



Hyperspectral Compressed Recovery

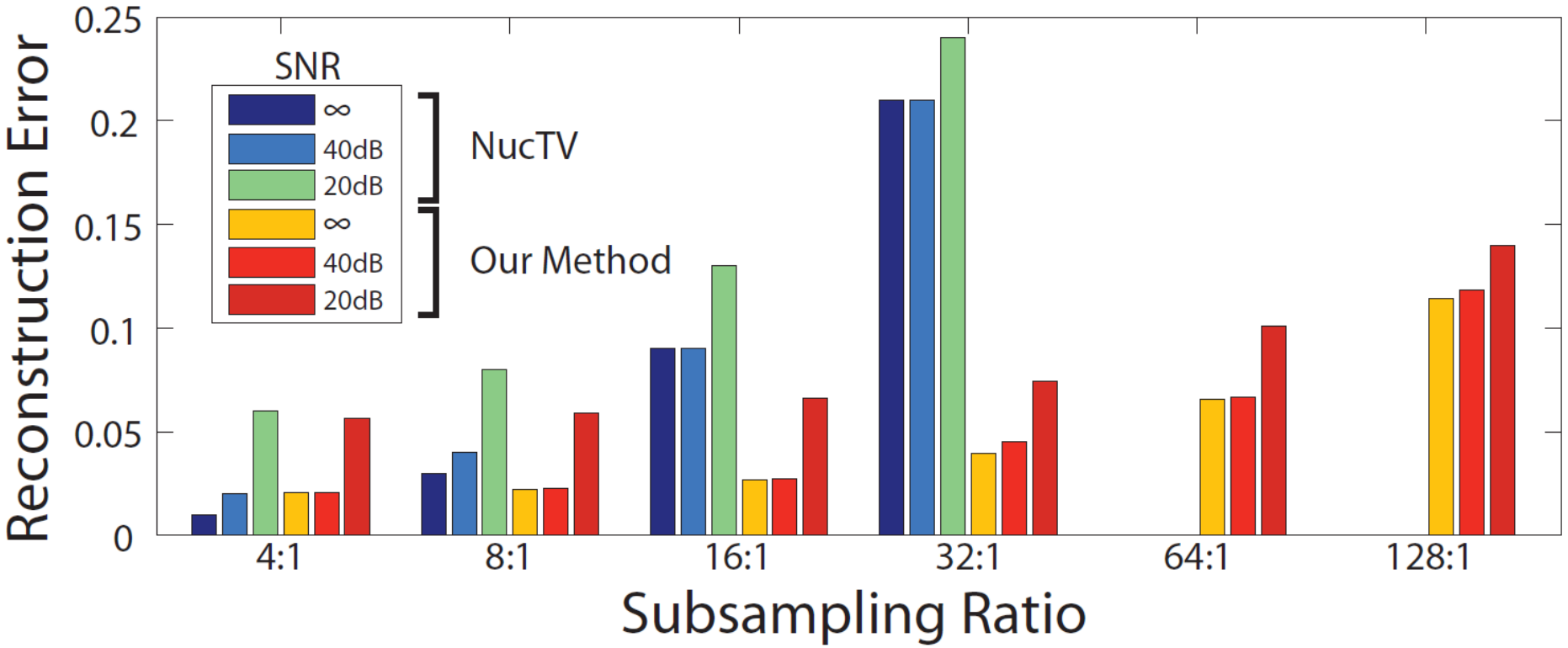
- Our method

$$\min_{U,V} \|Y - \Phi(UV^\top)\|_F^2 + \lambda \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v$$

- (U,V) have 15 columns
- Problem size reduced by 91.6%
- Computation per Iteration
  - Calculate gradient
  - 15 TV Proximal Operators
- Random Initializations

Hyperspectral Compressed Recovery

$$\frac{\|X_{true} - UV^\top\|_F}{\|X_{true}\|_F}$$



Conclussions

- Structured Low Rank Matrix Factorization
  - Structure on the factors captured by the Projective Tensor Norm
  - Efficient optimization for Large Scale Problems

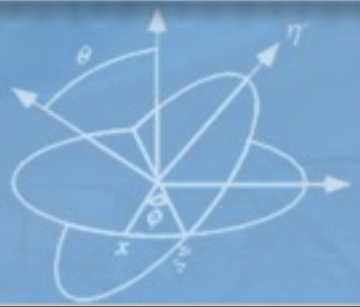
- Local minima of the non-convex factorized form are global minima of both the convex and non-convex forms

- Advantages in Applications
  - Neural calcium image segmentation
  - Compressed recovery of hyperspectral images

JHU vision lab

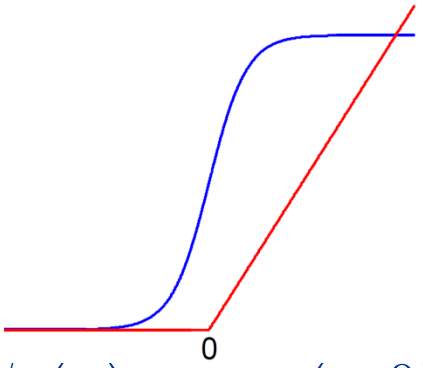
Global Optimality in Positively Homogeneous Factorization

René Vidal
Center for Imaging Science
Institute for Computational Medicine

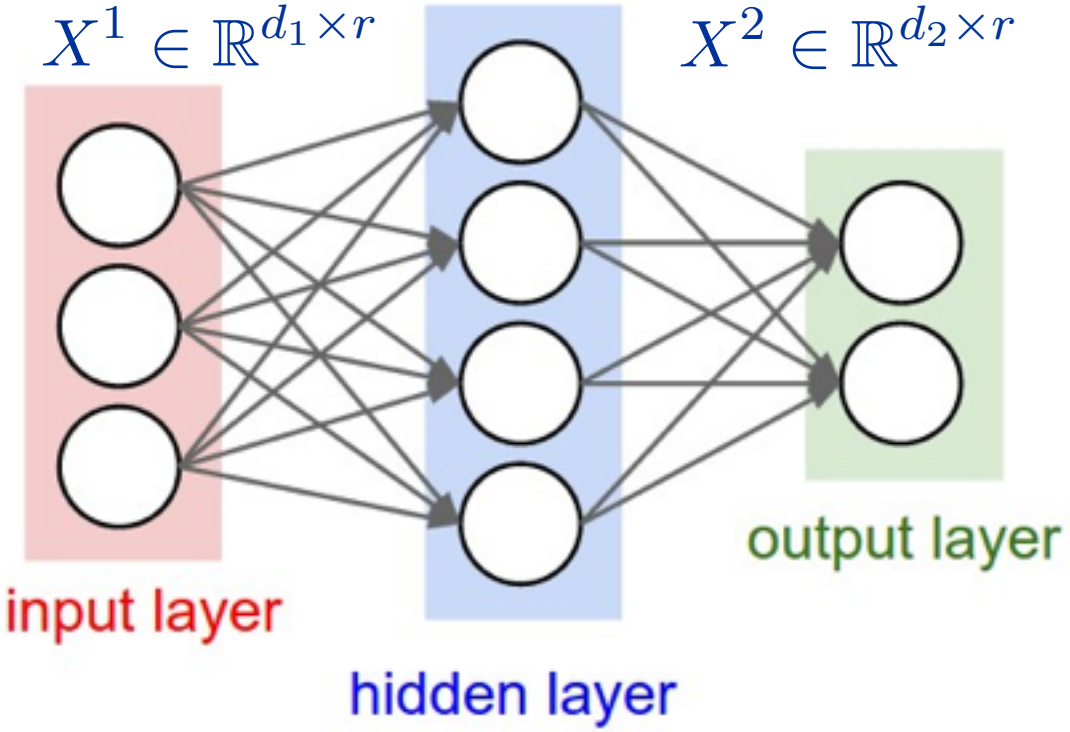


From Matrix Factorizations to Deep Learning

- Two-layer NN
  - Input: $V \in \mathbb{R}^{N \times d_1}$
  - Weights: $X^k \in \mathbb{R}^{d_k \times r}$
  - Nonlinearity: ReLU



$$\psi_1(x) = \max(x, 0)$$



- "Almost" like matrix factorization
  - r = rank
  - r = #neurons in hidden layer

$$\Phi(X^1, X^2) = \psi_1(VX^1)(X^2)^\top$$

From Matrix Factorizations to Deep Learning

- Recall the generalized factorization problem

$$\min_{X^1,\ldots,X^K} \ell(Y, \Phi(X^1,\ldots,X^K)) + \lambda\Theta(X^1,\ldots,X^K)$$

- Matrix factorization is a particular case where K=2

$$\Phi(U,V) = \sum_{i=1}^{r} U_i V_i^\top, \;\; \Theta(U,V) = \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v$$

- Both Φ and Θ are sums of positively homogeneous functions

$$f(\alpha X^1,\ldots,\alpha X^K) = \alpha^p f(X^1,\ldots,X^K) \quad \forall \alpha \geq 0$$

- Other examples
  - ReLU + max pooling is positively homogeneous of degree 1

“Matrix Multiplication” for K > 2

- In matrix factorization we have

$$\Phi(U, V) = UV^\top = \sum_{i=1}^{r} U_i V_i^\top$$

- By analogy we define

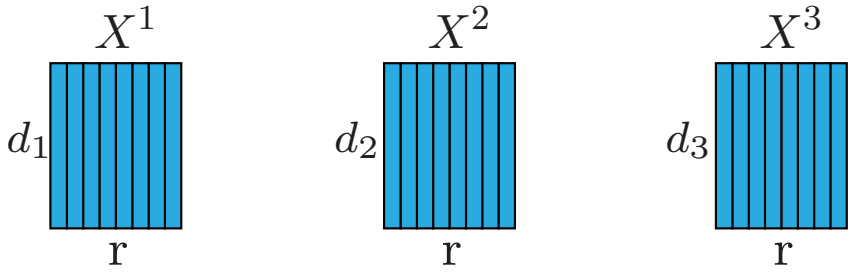
$$\Phi(X^1, \ldots, X^K) = \sum_{i=1}^{r} \phi(X_i^1, \ldots, X_i^K)$$

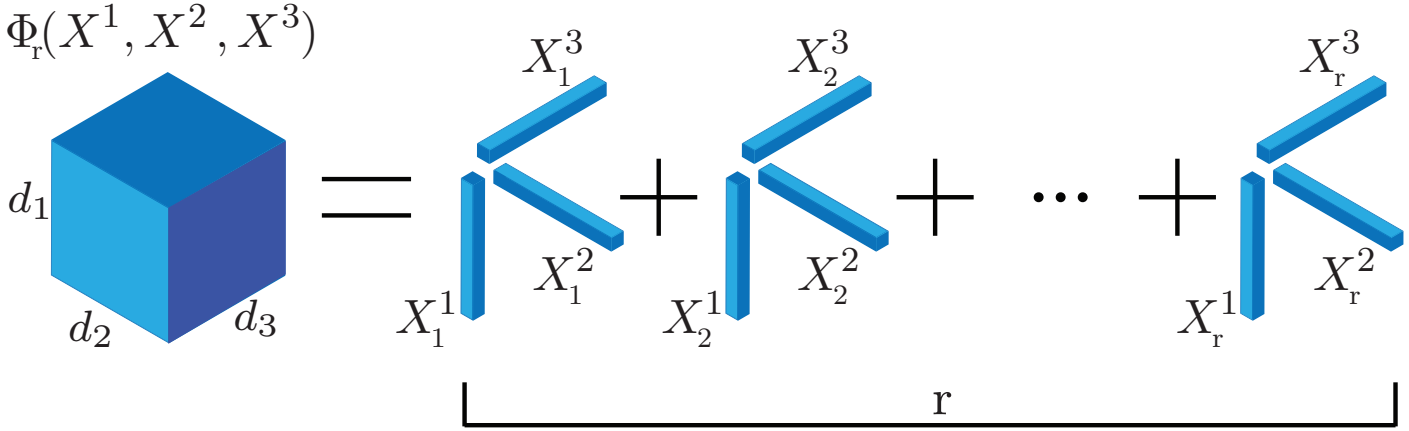
where X^k is a tensor, X_i^k is its i-th slice along its last dimension, and ϕ is a positively homogeneous function

- Examples
  - Matrix multiplication: $\phi(X^1, X^2) = X^1 X^{2\top}$
  - Tensor product: $\phi(X^1, \ldots, X^K) = X^1 \otimes \cdots \otimes X^K$
  - ReLU neural network: $\phi(X^1, \ldots, X^K) = \psi_K(\cdots \psi_2(\psi_1(VX^1)X^2)\cdots X^K)$

Example: Tensor Factorization

$$\Phi(X^1, \ldots, X^K) = \sum_{i=1}^{r} \phi(X_i^1, \ldots, X_i^K)$$

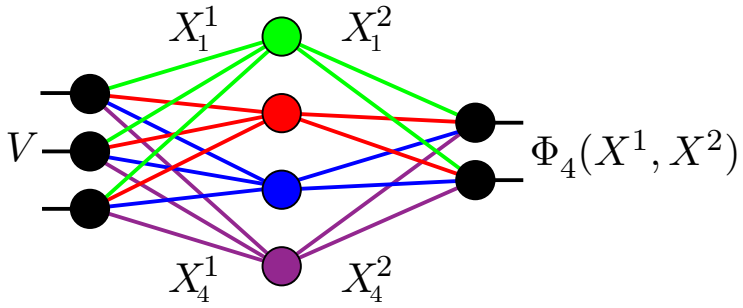




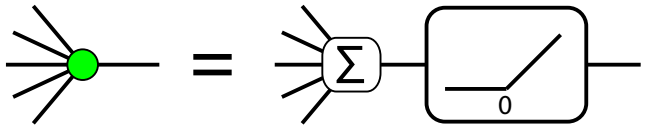
Example: Deep Learning

$$\Phi(X^1,\ldots,X^K) = \sum_{i=1}^{r} \phi(X_i^1,\ldots,X_i^K)$$

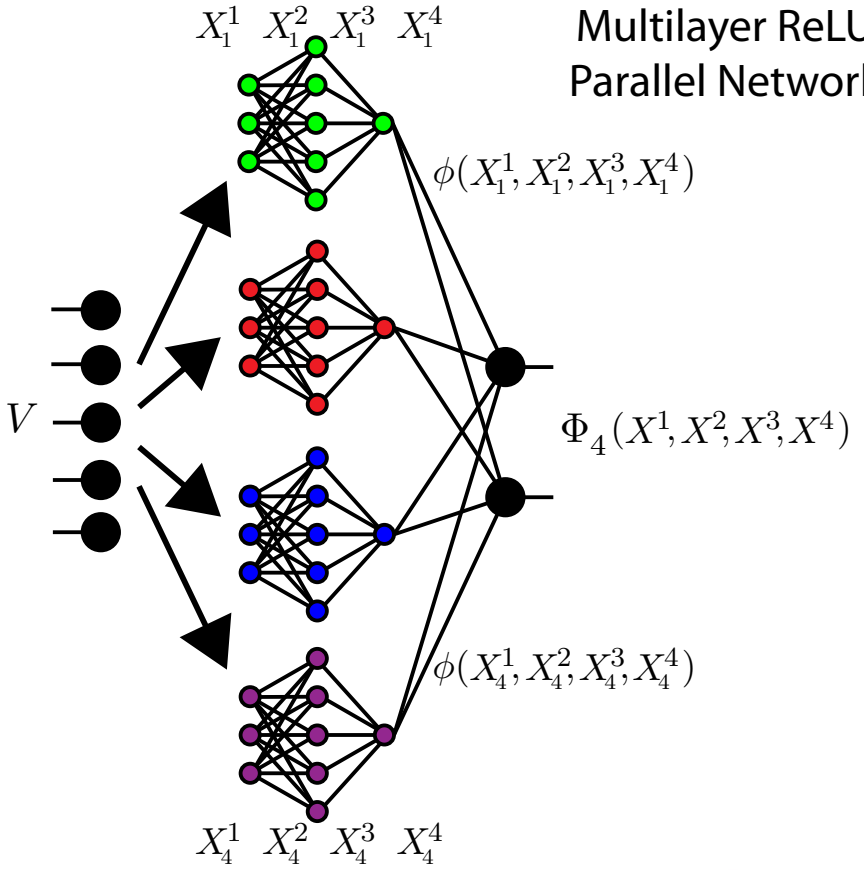
ReLU Network with One Hidden Layer



Rectified Linear Unit (ReLU)



Multilayer ReLU Parallel Network



“Projective Tensor Norm” for K > 2

- In matrix factorization we had

$$\|X\|_{u,v} = \min_{U,V} \sum_{i=1}^{r} \|U_i\|_u \|V_i\|_v \quad \text{s.t.} \quad UV^\top = X$$

- By analogy we define

$$\Omega_{\phi,\theta}(X) = \min_{\{X^k\}} \sum_{i=1}^{r} \theta(X_i^1, \ldots, X_i^K) \text{ s.t. } \Phi(X^1, \ldots, X^K) = X$$

  where θ is positively homogeneous of the same degree as ϕ

- **Proposition:** $\Omega_{\phi,\theta}$ is convex

Main Results

- **Theorem 1:** A local minimizer of the factorized formulation

$$\min_{\{X^k\}} \ell\Big(Y, \sum_{i=1}^{r} \phi(X_i^1, \dots, X_i^K)\Big) + \lambda \sum_{i=1}^{r} \theta(X_i^1, \dots, X_i^K)$$

such that for some i and all k $X_i^k = 0$ is a global minimizer for both the factorized problem and of the convex formulation

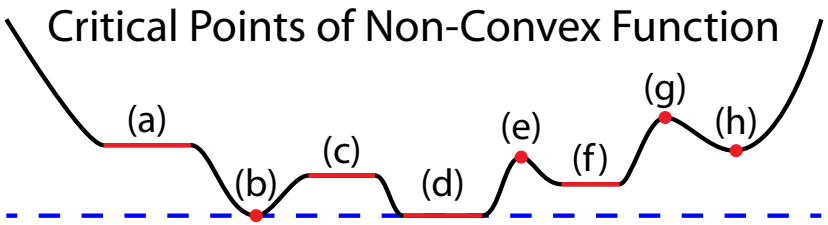
$$\min_{X} \ell(Y, X) + \lambda \Omega_{\phi,\theta}(X)$$

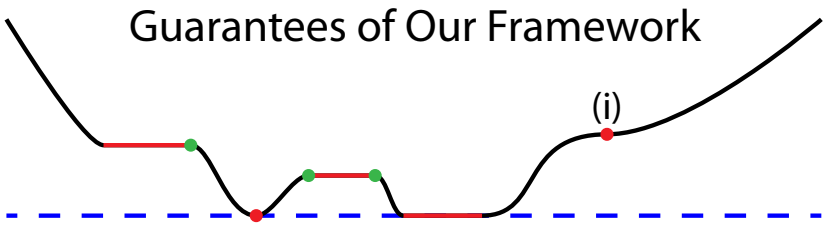
- Examples
  - Matrix factorization
  - Tensor factorization
  - Deep learning

[1] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv '15

Main Results

- **Theorem 2:** If the size of the network is large enough, local descent can reach a global minimizer from any initialization





- **Meta-Algorithm:**
  - If not at a local minima, perform local descent to reach a local minima
  - If optimality condition is satisfied, then local minima is global
  - If condition fails, choose descent direction, increase r <- r+1, and move along descent direction

[1] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv '15

Conclusions

- For many non-convex factorization problems, such as matrix factorization, tensor factorization, and deep learning, a local minimizer for the factors gives a global minimizer
- For matrix factorization, this
  - allows one to incorporate structure on the factors, and
  - gives efficient optimization method suitable for large problems
- For deep learning, this provides theoretical insights on why
  - many local minima give similar objective values
  - ReLU works better than sigmoidal functions
- While alternating minimization is efficient and guaranteed to converge, it is not guaranteed to converge to a local minimum

More Information,

Vision Lab @ Johns Hopkins University

http://www.vision.jhu.edu

Center for Imaging Science @ Johns Hopkins University

http://www.cis.jhu.edu

Thank You!