JHU vision lab

Global Optimality in Matrix and Tensor Factorization, Deep Learning & Beyond

Ben Haeffele and René Vidal
Center for Imaging Science
Johns Hopkins University

THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins

Center for Imaging Science

Impact of Deep Learning in Computer Vision

- 2012-2014 classification results in ImageNet

CNN
non-CNN

2012 Teams	%error
Supervision (Toronto)	15.3
ISI (Tokyo)	26.1
VGG (Oxford)	26.9
XRCE/INRIA	27.0
UvA (Amsterdam)	29.6
INRIA/LEAR	33.4

2013 Teams	%error
Clarifai (NYU spinoff)	11.7
NUS (singapore)	12.9
Zeiler-Fergus (NYU)	13.5
A. Howard	13.5
OverFeat (NYU)	14.1
UvA (Amsterdam)	14.2
Adobe	15.2
VGG (Oxford)	15.2
VGG (Oxford)	23.0

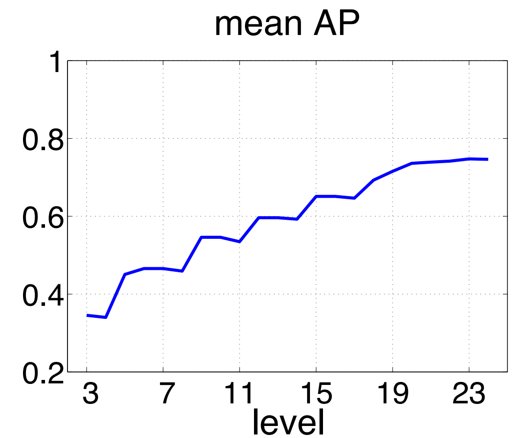
2014 Teams	%error
GoogLeNet	6.6
VGG (Oxford)	7.3
MSRA	8.0
A. Howard	8.1
DeeperVision	9.5
NUS-BST	9.7
TTIC-ECP	10.2
XYZ	11.2
UvA	12.1

- 2015 results: MSR under 3.5% error using 150 layers!

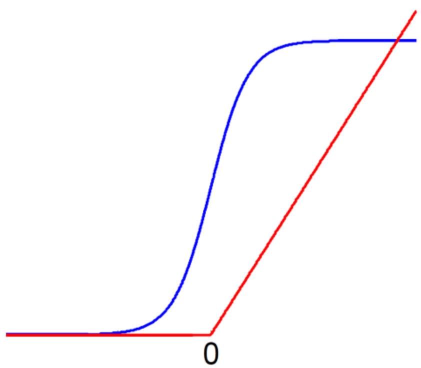
Slide from Yann LeCun's CVPR'15 plenary and ICCV'15 tutorial intro by Joan Bruna

Why These Improvements in Performance?

- Features are learned rather than hand-crafted
- More layers capture more invariances [1]



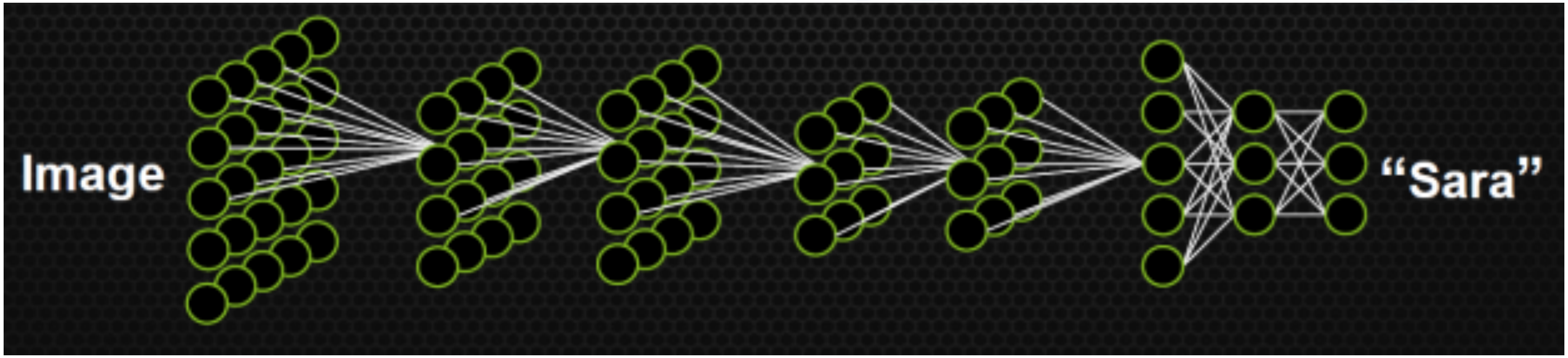
- More data to train deeper networks
- More computing (GPUs)
- Better regularization: Dropout
- New nonlinearities
  - Max pooling, Rectified linear units (ReLU)

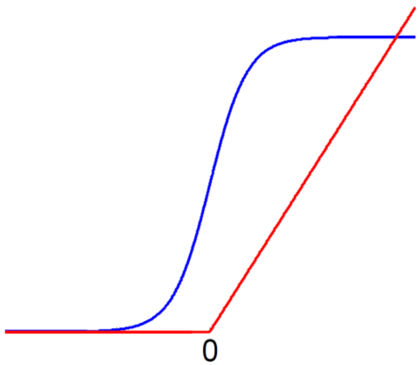


- Theoretical understanding of deep networks remains shallow

[1] Razavian, Azizpour, Sullivan, Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition. CVPRW'14.

Deep Learning Problem is Non Convex





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

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

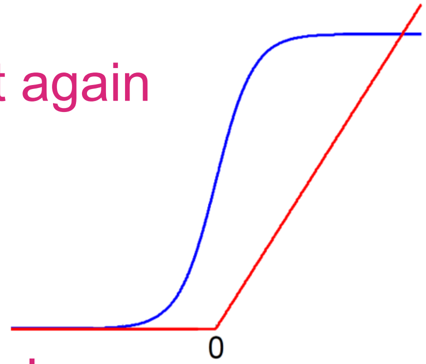
How is Non Convexity Handled?

- 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)

Prior Work on Optimization for Neural Nets

- **Earlier work**
  - No spurious local optima for linear networks (Baldi & Hornik ’89)
  - Stuck in local minima (Brady ’89, Gori & Tesi ’92), but guaranteed to converge for linearly separable data (Gori & Tesi ’92)
  - Manifold of spurious local optima (Frasconi ’97)
- **Recent work**
  - Convex neural networks in infinite number of variables: Bengio ‘05
  - Networks with many hidden units can learn polynomials: Andoni‘14
  - The loss surface of multilayer networks: Choromanska ’15
  - Attacking the saddle point problem: Dauphin ‘14
  - Effect of gradient noise on the energy landscape: Chaudhuri ‘15
  - Guaranteed training of NNs using tensor methods: Janzamin ’15
- **Today**
  - Guarantees of global optimality in neural network training: Haeffele ‘15

Main Results

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

- **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

Benjamin D. Haeffele, Rene Vidal. Global Optimality in Tensor Factorization, Deep Learning, and Beyond. arXiv:1506.07540, 2015

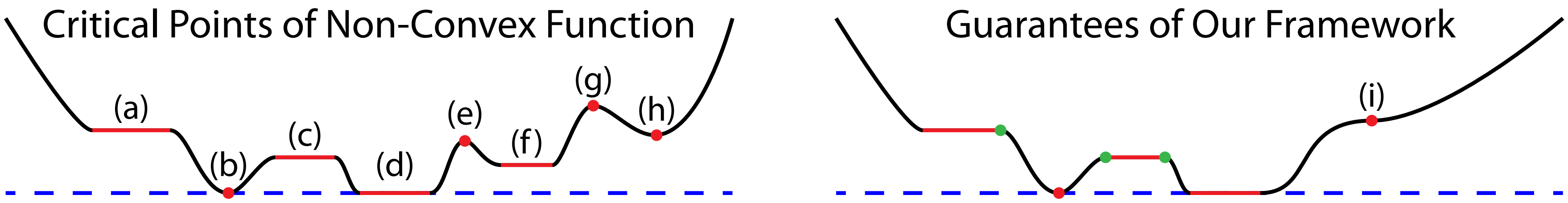
Main Results

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

- **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:** spurious local minima guaranteed not to exist

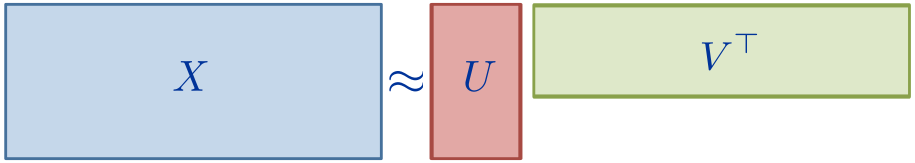


Benjamin D. Haeffele, Rene Vidal. Global Optimality in Tensor Factorization, Deep Learning, and Beyond. arXiv:1506.07540, 2015

Outline

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

- 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

Center for IMAGING SCIENCE

JHU vision lab

Global Optimality in Structured Matrix Factorization

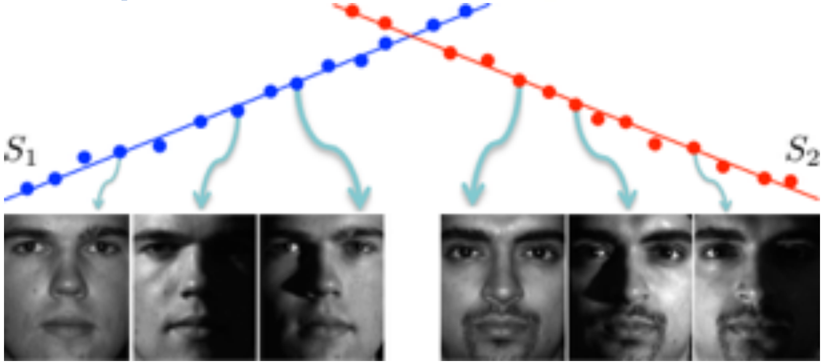
Ben Haeffele and René Vidal
Center for Imaging Science
Johns Hopkins University

THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins

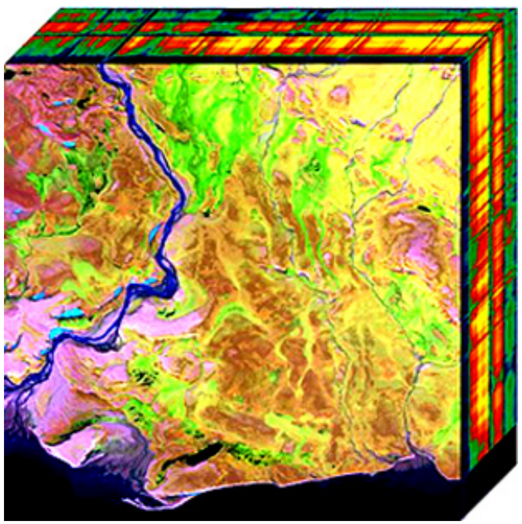
Center for Imaging Science

Low-Rank Modeling

- Models involving factorization are ubiquitous
  - Principal Component Analysis
  - Nonnegative Matrix Factorization
  - Sparse Dictionary Learning
  - Low-Rank Matrix Completion
  - Robust PCA



Face clustering and classification



Hyperspectral imaging

NETFLIX

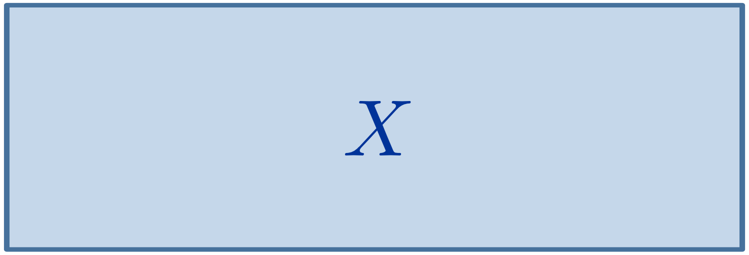
Recommendation systems

Affine structure from motion

Typical Low-Rank Formulations

- **Convex formulations**:

$$\min_{X} \ell(Y, X) + \lambda\Theta(X)$$



  - Low-rank matrix approximation
  - Low-rank matrix completion
  - Robust PCA

✓ Convex

✳ Large problem size

✳ Unstructured factors

- **Factorized formulations**:

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

  - Principal component analysis
  - Nonnegative matrix factorization
  - Sparse dictionary learning

✳ Non-Convex

✓ Small problem size

✓ Structured factors

Convex Formulations of Matrix Factorization

- **Convex formulations**:
  - ℓ, Θ : convex in X

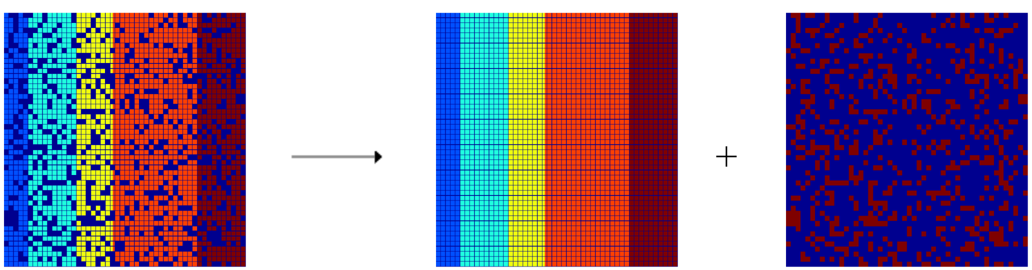
$$\min_X \; \ell(Y, X) + \lambda\, \Theta(X)$$

- Low-rank matrix approximation:

$$\min_X \; \frac{1}{2}\|Y - X\|_F^2 + \lambda\, \|X\|_* \qquad \|X\|_* = \sum \sigma_i(X)$$

- Robust PCA:

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



✓ Convex

✳ Large problem size

✳ Unstructured factors

EJ Candès, B Recht. Exact matrix completion via convex optimization. Foundations of Computational mathematics, 2009.
RH Keshavan, A Montanari, S Oh. Matrix completion from a few entries. IEEE Trans. on Information Theory, 2010.
EJ Candès, T Tao. The power of convex relaxation: Near-optimal matrix completion. IEEE Trans. on Information Theory, 2010
Candes, Li, Ma, Wright. Robust Principal Component Analysis? Journal of the ACM, 2011.
H Xu, C Caramanis, S Sanghavi. Robust PCA via outlier pursuit. NIPS 2010

Center for IMAGING SCIENCE

Factorized Formulations Matrix Factorization

- **Factorized formulations**:
  - $\ell(Y, X)$: convex in X

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

- PCA [1]: $\min_{U,V} \|Y - UV^\top\|_F^2 \quad \text{s.t.} \quad U^\top U = I$
- NMF [2]: $\min_{U,V} \|Y - UV^\top\|_F^2 \quad \text{s.t.} \quad U \geq 0, V \geq 0$
- SDL [3-5]: $\min_{U,V} \|Y - UV^\top\|_F^2 \quad \text{s.t.} \quad \|U_i\|_2 \leq 1, \|V_i\|_0 \leq r$

✓ Small problem size

✓ Structured factors

✳ Need to specify size a priori

✳ Non-convex optimization problem

[1] Jolliffe. Principal component analysis. Springer, 1986
[2] Lee and Seung. "Learning the parts of objects by non−negative matrix factorization." Nature, 1999
[3] Olshausen and Field, “Sparse coding with an overcomplete basis set: A strategy employed by v1?,” Vision Research, 1997
[4] Engan, Aase, and Hakon-Husoy, “Method of optimal directions for frame design,” ICASSP 1999
[5] Aharon, Elad, Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation", TSP 2006

Main Results

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

- **Assumptions:**
  - $\ell(Y, X)$: convex and once differentiable in X
  - Θ : sum of positively homogeneous functions of degree 2

$$\Theta(U, V) = \sum_{i=1}^{r} \theta(U_i, V_i), \quad \theta(\alpha u, \alpha v) = \alpha^2 \theta(u, v), \forall \alpha \geq 0$$

- **Theorem 1:** A local minimizer (U,V) such that for some *i* $U_i = V_i = 0$ is a global minimizer

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

B. Haeffele, E. Young, R. Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing. ICML 2014
Benjamin D. Haeffele, Rene Vidal. Global Optimality in Tensor Factorization, Deep Learning, and Beyond. arXiv:1506.07540, 2015

Main Results: Nuclear Norm Case

- **Convex problem** **Factorized problem**

$$\min_{X} \ell(Y, X) + \lambda \|X\|_* \qquad \min_{U,V} \ell(Y, UV^\top) + \lambda \Theta(U, V)$$

- 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 1:** 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"

The following papers study the case of a square loss function using techniques from semi-definite programming:

[1] S. Burer and R. Monteiro. Local minima and convergence in low- rank semidefinite programming. Math. Prog., 103(3):427–444, 2005.
[2] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition," in IEEE International Conference on Computer Vision, 2013, pp. 2488–2495.

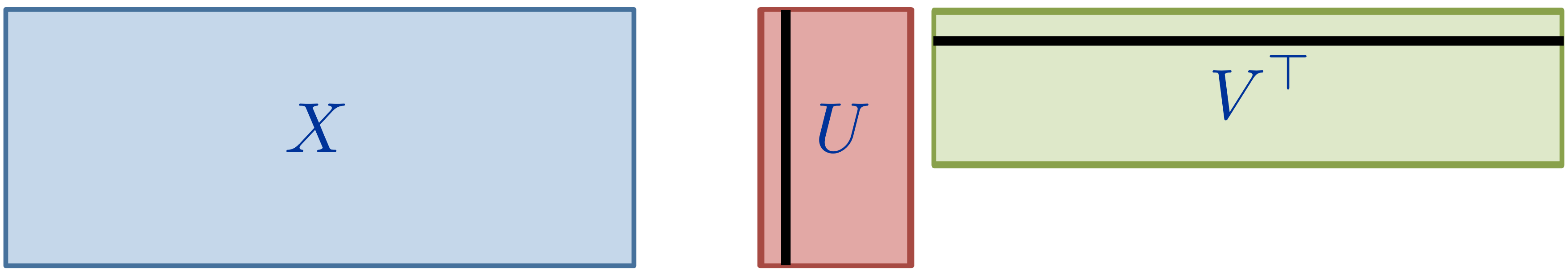
Main Results: Nuclear Norm Case

- **Convex problem**

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

Factorized problem

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



- **Theorem 1:** 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

The following papers study the case of a square loss function using techniques from semi-definite programming:
[1] S. Burer and R. Monteiro. Local minima and convergence in low- rank semidefinite programming. Math. Prog., 103(3):427–444, 2005.
[2] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition," in IEEE International Conference on Computer Vision, 2013, pp. 2488–2495.

Main Results: Projective 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

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

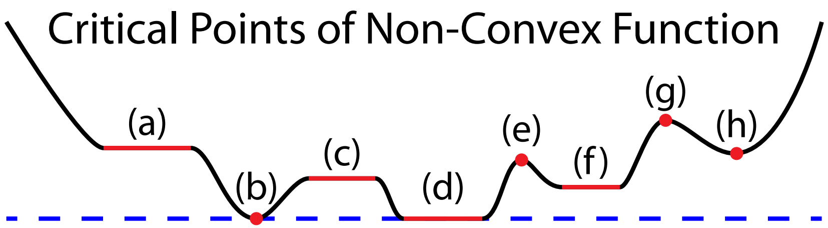
such that for some i $U_i = V_i = 0$, is a global minimizer of both the factorized problem and of the convex problem

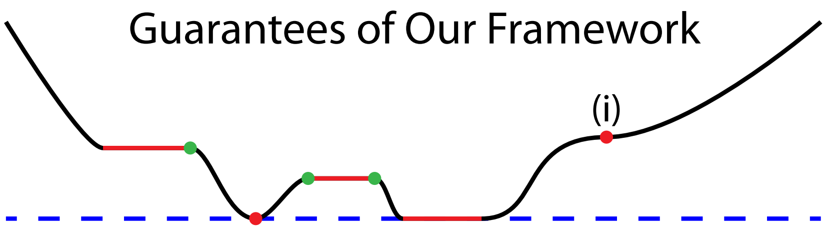
$$\min_{X} \ell(Y, X) + \lambda \|X\|_{u,v}$$

[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

Main Results: Projective 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
  - At local minima, test if Theorem 1 is satisfied. If yes => global minima
  - If not, increase size of factorization and find descent direction (u,v)

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

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

Algorithm: Projective Tensor Norm Case

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

- Convex in U given V and vice versa
- Alternating proximal gradient descent
  - Calculate gradient of smooth term
  - Compute proximal operator
  - Acceleration via extrapolation
- Advantages
  - Easy to implement
  - Highly parallelizable
  - Guaranteed to converge to Nash equilibrium (may not be local min) [1]

Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," SIAM Journal of Imaging Sciences, vol. 6, no. 3, pp. 1758–1789, 2013.

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

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

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

Center for IMAGING SCIENCE

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

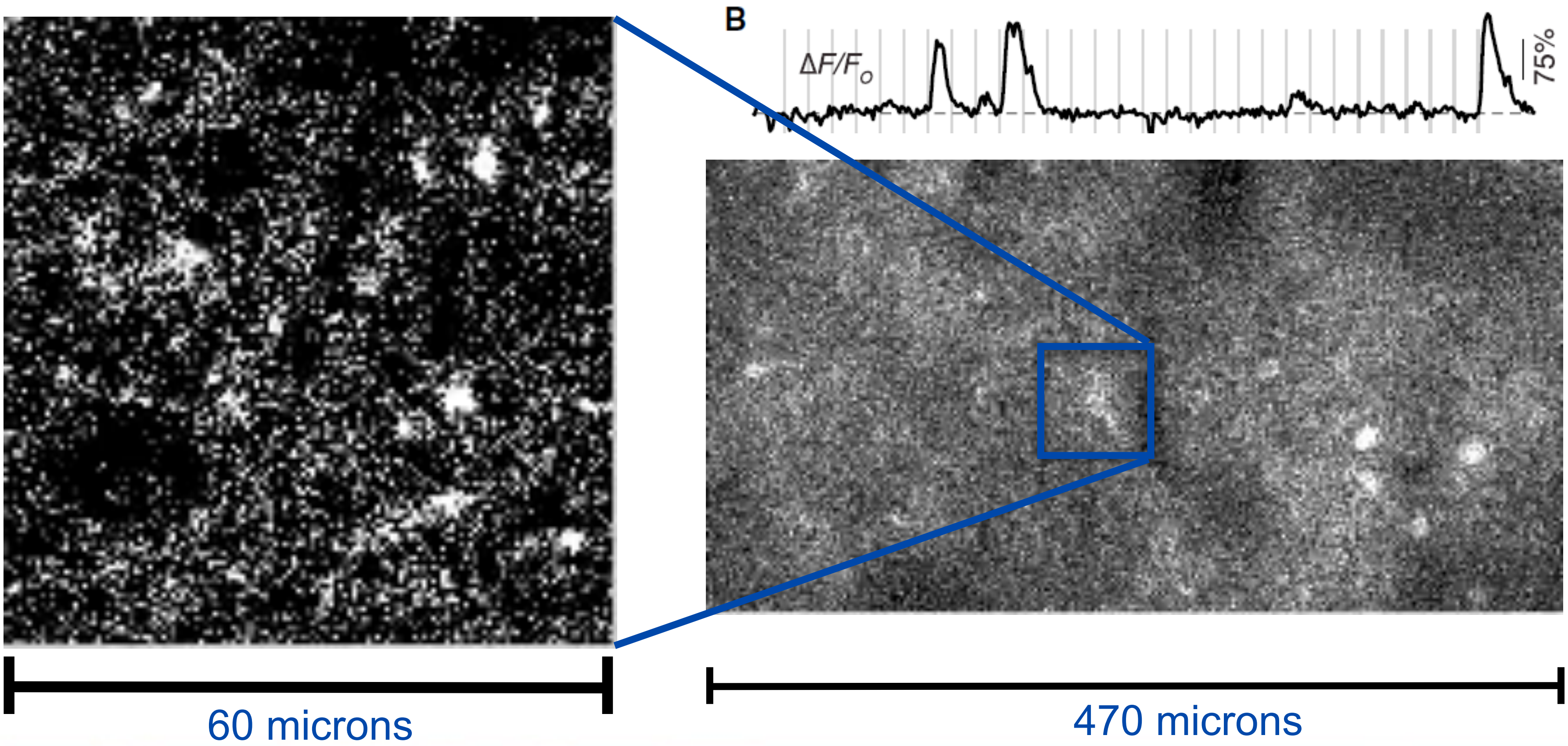
$$\min_{U,V} \|Y - UV^\top\|_1 + \lambda \sum_i |U_i|_2 |V_i|_2$$

- Not an example because loss is not differentiable

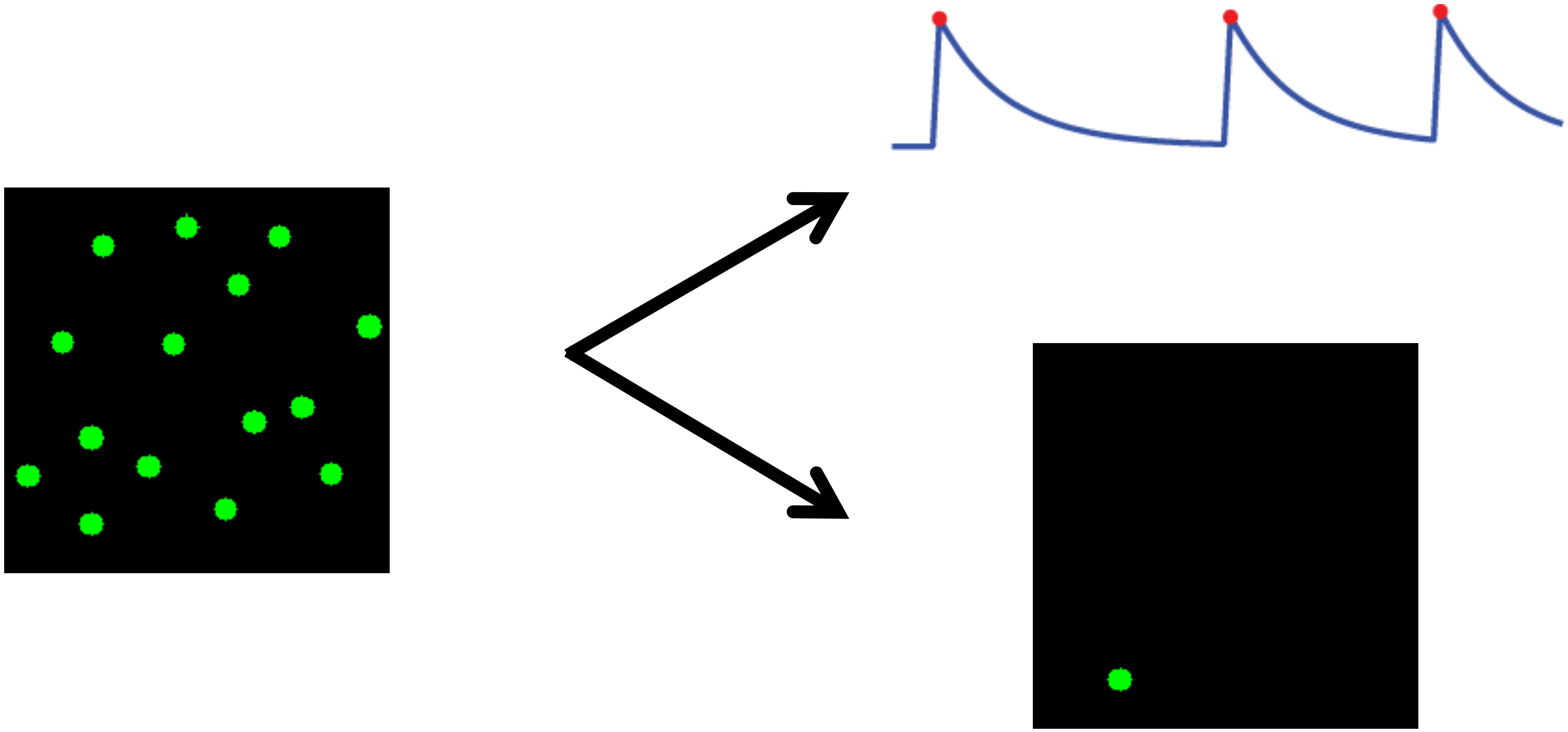
[1] Candes, Li, Ma, Wright. Robust Principal Component Analysis? Journal of the ACM, 2011.

Application: Calcium Imaging Segmentation

- Fluorescent microscopy technique
  - Optical recording of brain activity
  - Neurons “flash” when active electrically



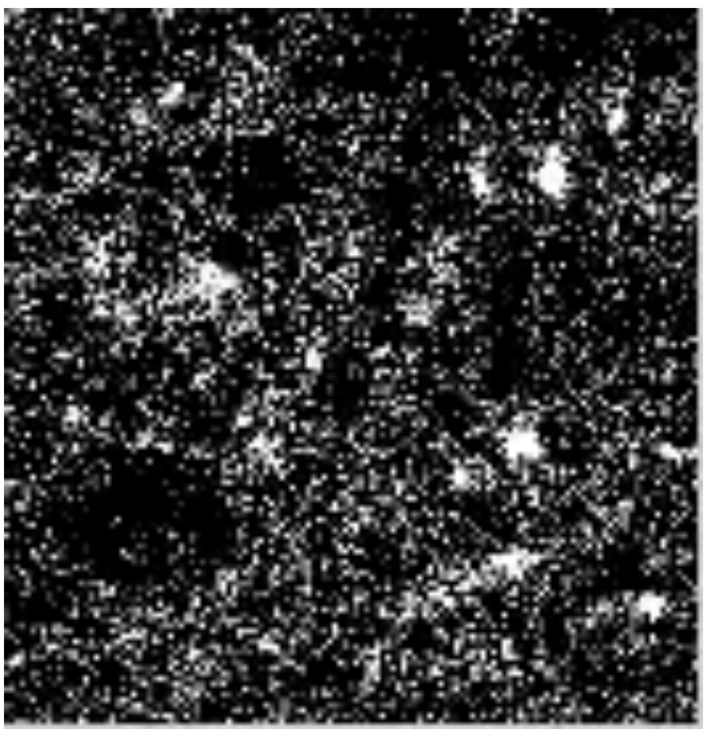
Application: Calcium Imaging Segmentation

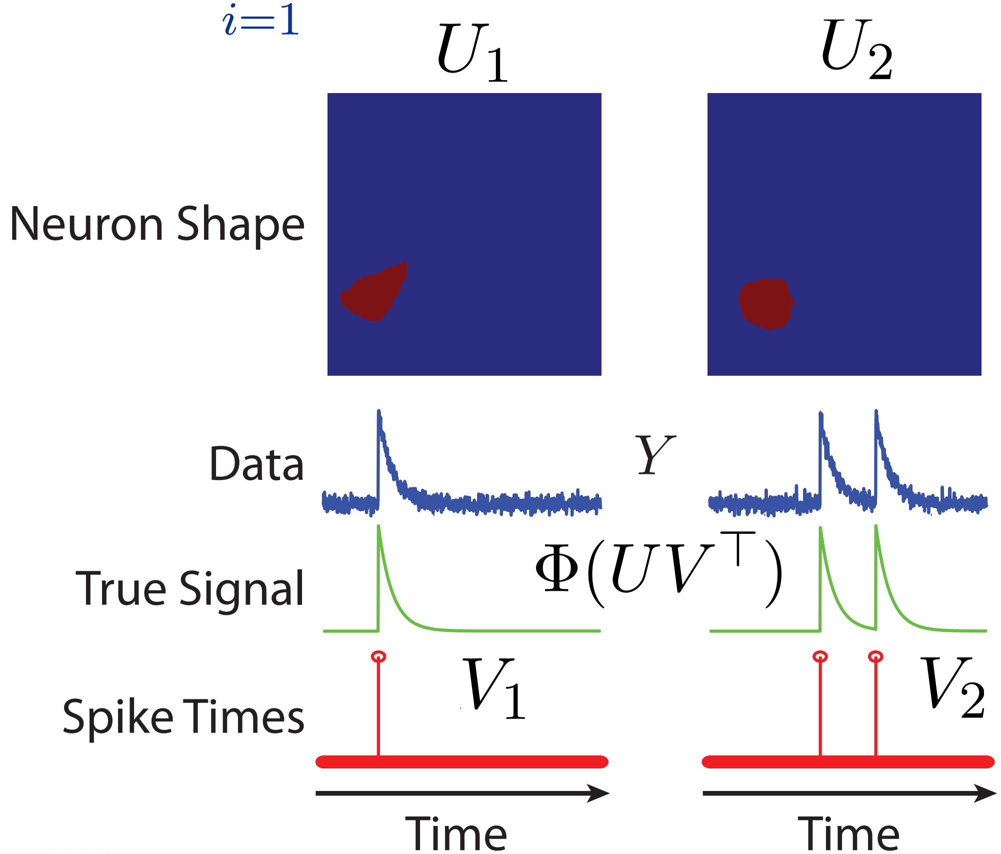


Application: Calcium Imaging Segmentation

- Find neuronal shapes and spike trains in calcium imaging

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



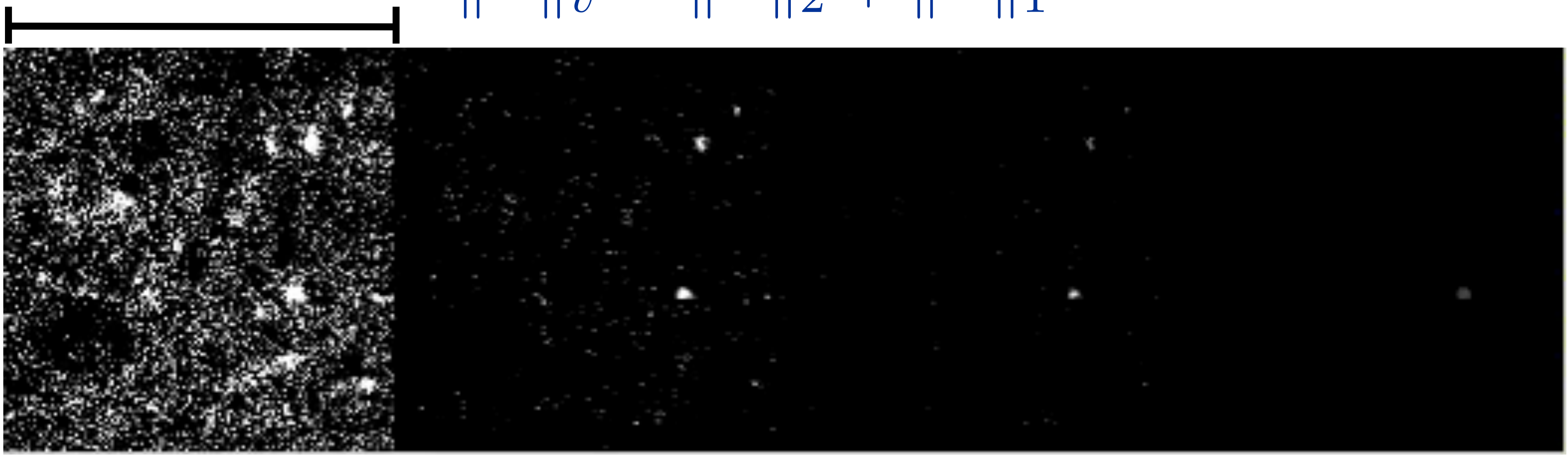


In Vivo Results (Small Area)

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

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

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



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

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

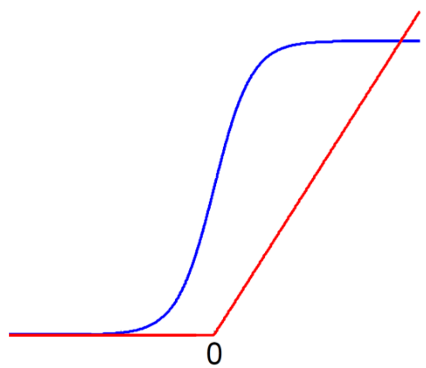
JHU vision lab

Global Optimality in Positively Homogeneous Factorization

Ben Haeffele and René Vidal
Center for Imaging Science
Johns Hopkins University

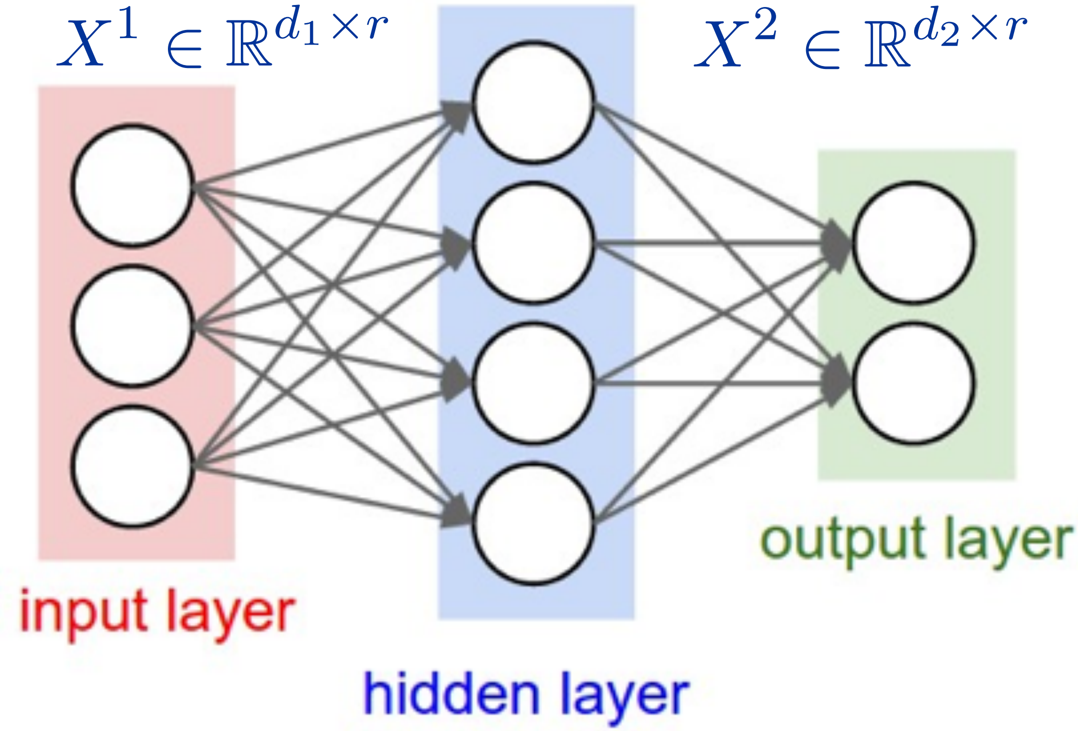
From Matrix Factorizations to Deep Learning

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



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

$X^1 \in \mathbb{R}^{d_1 \times r}$ $X^2 \in \mathbb{R}^{d_2 \times r}$



- “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

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

- 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

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

- 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

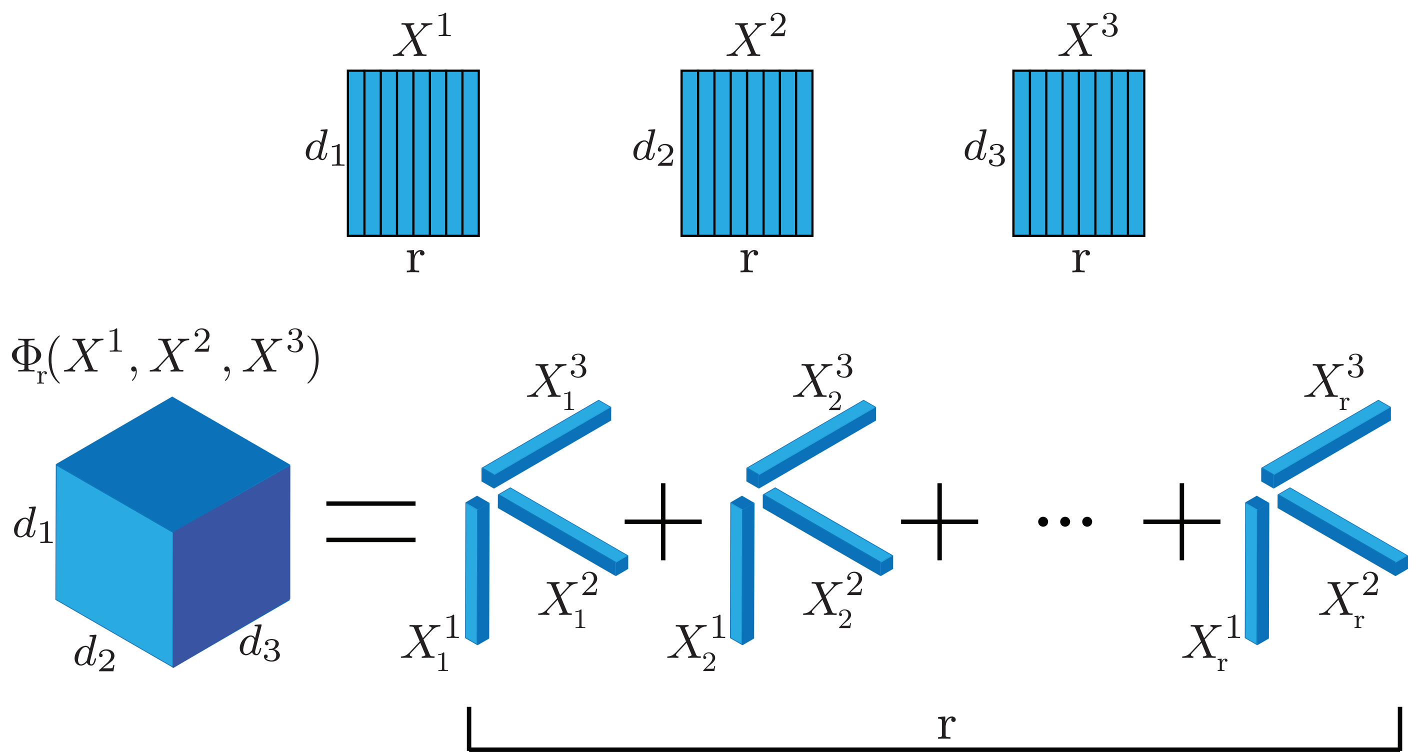
$$\Phi(X^1,\ldots,X^K) = \sum_{i=1}^{r} \phi(X_i^1,\ldots,X_i^K)$$

  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: CP Tensor Factorization

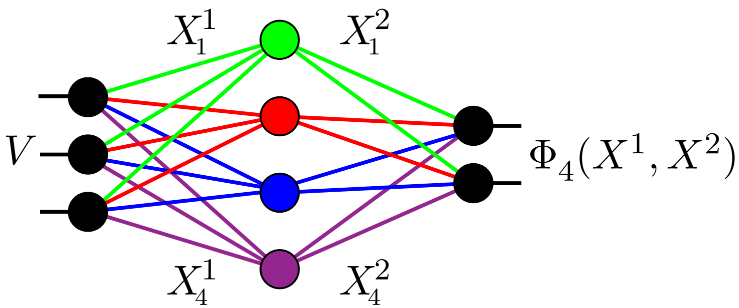
$$\Phi(X^1, \ldots, X^K) = \sum_{i=1}^{r} \phi(X_i^1, \ldots, X_i^K)$$



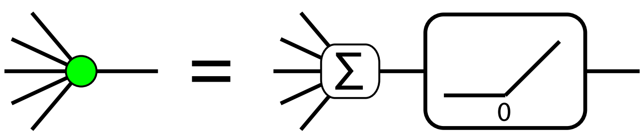
Example: Deep Learning

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

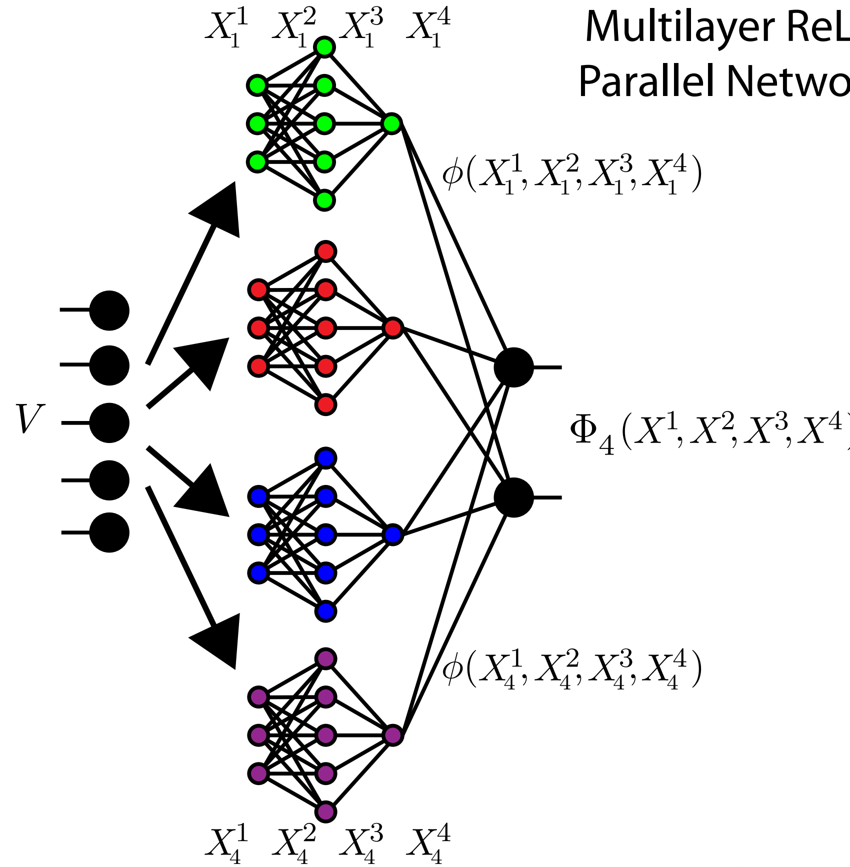
ReLU Network with One Hidden Layer



Rectified Linear Unit (ReLU)



Multilayer ReLU Parallel Network



Factorization Regularization for “K > 2”

- In matrix factorization we had “generalized nuclear norm”

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

- By analogy we define “nuclear deep net regularizer”

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

where θ is positively homogeneous of the same degree as ϕ

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

- **Intuition:** regularizer Θ “comes from a convex function”

Examples of Deep Network Regularizers

- Different norms for different properties on each factor

$$\theta(X_i^1, \ldots, X_i^K) = \prod_{k=1}^{K} \|X_i^k\|_{(k)}$$

- Different norms plus conic set constraints on the factors

$$\theta(X_i^1, \ldots, X_i^K) = \prod_{k=1}^{K} \left(\|X_i^k\|_{(k)} + \delta_{C_k}(X_i^k) \right) \qquad \delta_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$$

- Conic set examples
  - Kernel of linear operator — $\{x : Ax = 0\}$
  - Inequalities w.r.t. linear operator — $\{x : Ax \geq 0\}$
  - Constraints on non−zero support — $\{x : \|x\|_0 \leq n\}$
  - Semidefinite matrices — $\{x : x \in S_+^n\}$

Main Results

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

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

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

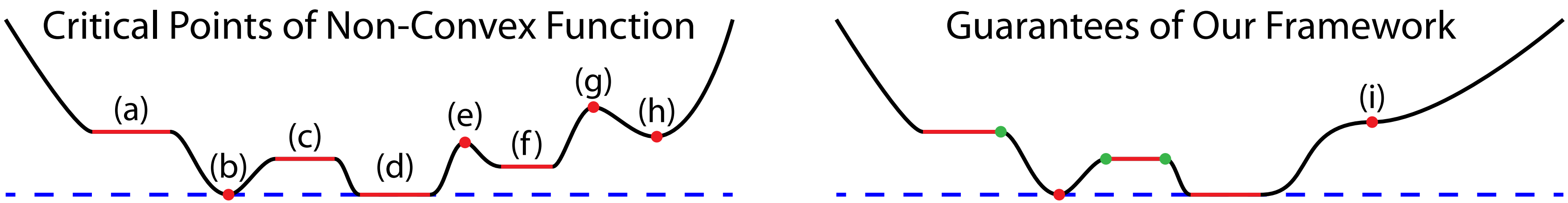
$$\min_X \ell(Y, X) + \lambda \Omega_{\phi,\theta}(X)$$

- 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
  - At a local minima, test if Theorem 1 is satisfied. If yes => global minima
  - If not, increase size by 1 (add network in parallell) and continue
  - Maximum r guaranteed to be bounded by the dimensions of the network output

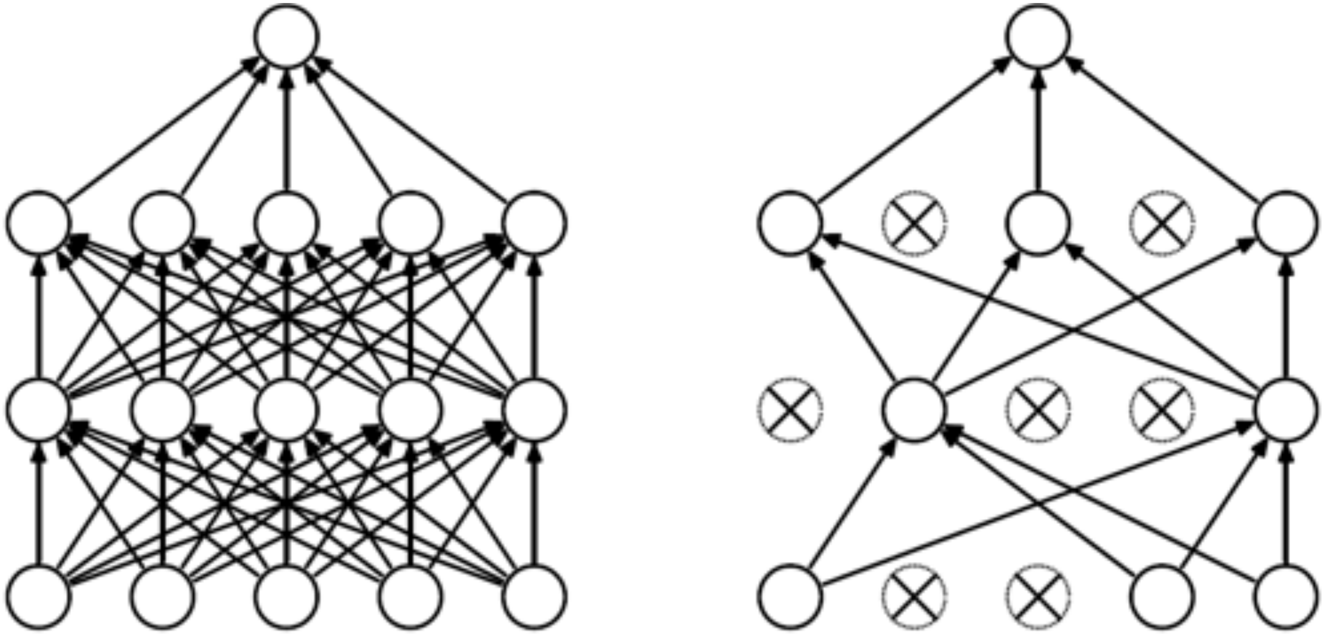
[1] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv '15

Current Limitations

- Requires networks with parallel architecture
  - Future work to explore more general regularization strategies to control other aspects of the network architecture
- Results only apply to local minima, not saddle points
  - Finding descent direction from saddle point can be NP−Hard
- Upper bound on size of network is impractically large
  - O(# of training examples in dataset)
  - But, this is a worst case upper bound for any possible initialization

Relation to Dropout

- Our theory suggests that a highly parallel architecture is advantageous for optimization
- Similar to dropout regularization (not an exact analogy)
  - Sum of exponential number of subnetworks



(a) Standard Neural Net (b) After applying dropout.

[1] Srivastava, et al, "Dropout: a simple way to prevent neural networks from overfitting." Journal of Machine Learning Research, 2014.

Balanced Degrees of Homogeneity

- Weight decay is often cited as not performing as well as dropout in ReLU networks [1–3].
  - Ex: L2 decay

$$\min_{X^1,\ldots,X^K} \ell(Y, \Phi(X^1,\ldots,X^K)) + \lambda \sum_{k=1}^{K} \|X^k\|_F^2$$

- Degrees of homogeneity are not typically balanced

$$\Phi(\alpha X^1,\ldots,\alpha X^K) = \alpha^K \Phi(X^1,\ldots,X^K)$$

$$\sum_{k=1}^{K} \|\alpha X^k\|_F^2 = \alpha^2 \sum_{k=1}^{K} \|X^k\|_F^2$$

- Proposition: If K > 2 there exist spurious local minima

[1] Srivastava, et al, "Dropout: a simple way to prevent neural networks from overfitting." JMLR, 2014.
[2] Krizhevsky, et al, “Imagenet classification with deep convolutional neural networks.” NIPS, 2012.
[3] Wan et al, “Regularization of neural networks using dropconnect.” ICML, 2013.

Conclusions and Future Directions

- **Size matters**
  - Optimize not only the network weights, but also the network size
  - Today: size = number of neurons or number of parallel networks
  - Tomorrow: size = number of layers + number of neurons per layer
- **Regularization matters**
  - Use “positively homogeneous regularizer” of same degree as network
  - How to build a regularizer that controls number of layers + number of neurons per layer
- **Not done yet**
  - Checking if we are at a local minimum or finding a descent direction can be NP hard
  - Need “computationally tractable” regularizers

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!