JHU vision lab

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

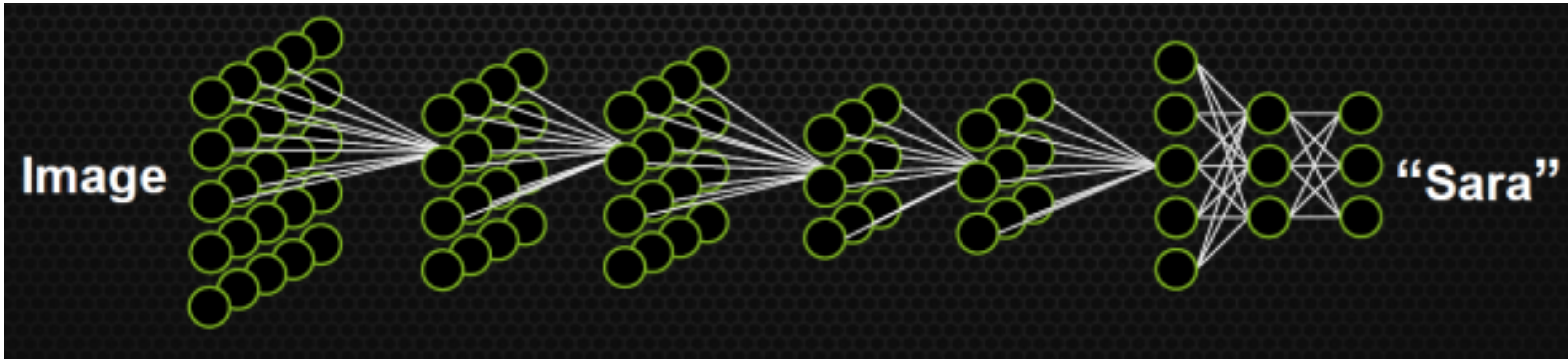
Ben Haeffele and René Vidal
Center for Imaging Science
Johns Hopkins University

THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins

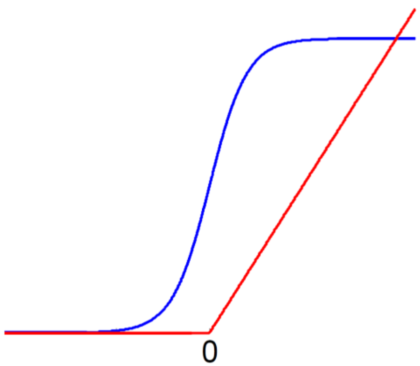
Center for Imaging Science

Learning Problem for Neural Networks

- The learning problem is non-convex



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



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

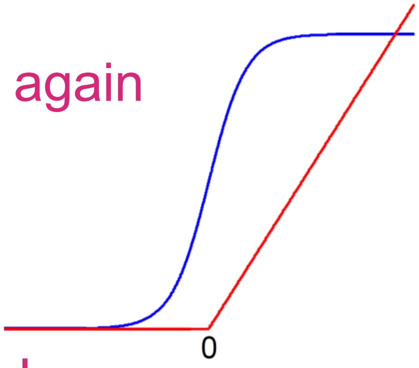
How is Non Convexity Handled?

- The learning problem is non-convex

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

  – Back-propagation, alternating minimization, descent method

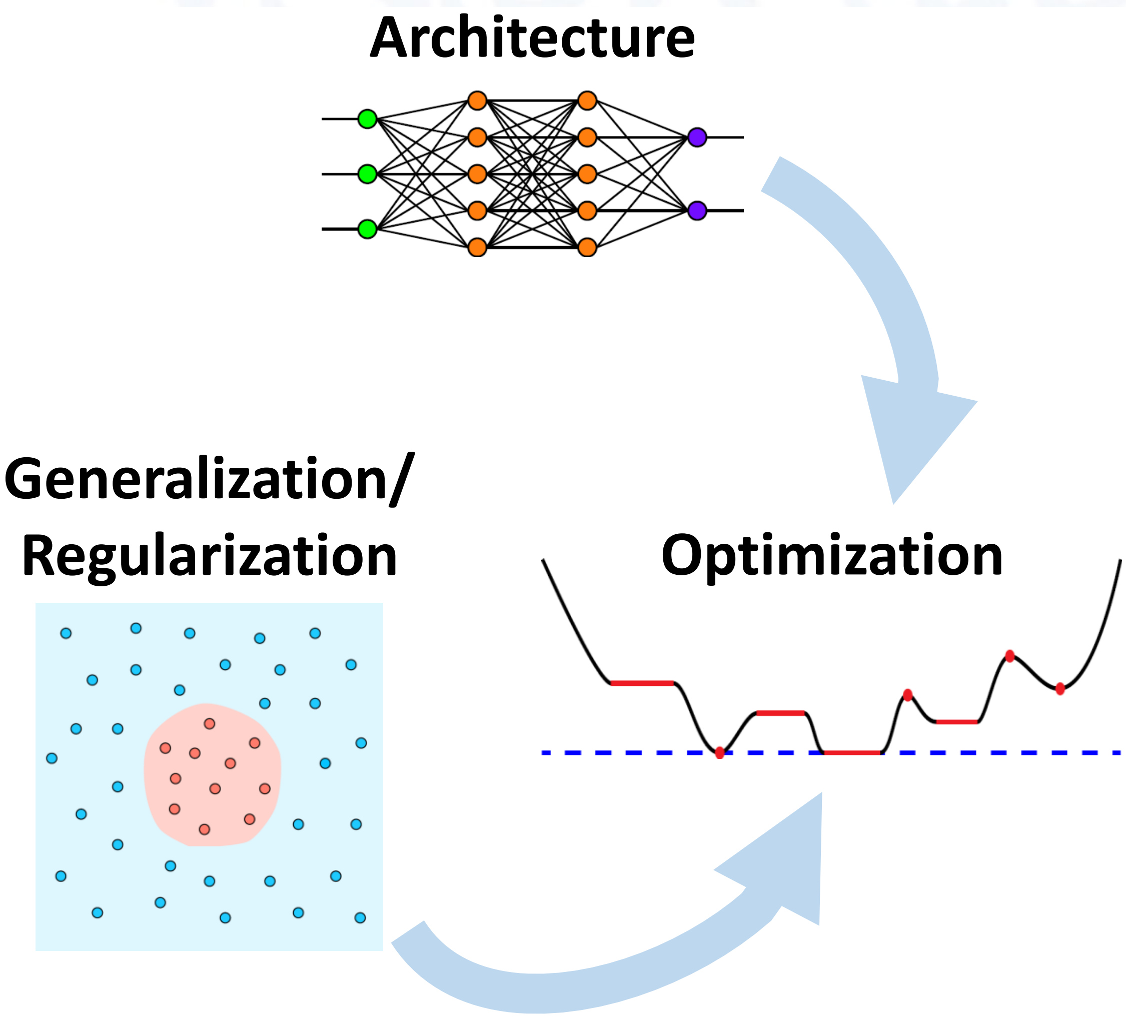
- To get a good local minima
  – Random initialization
  – If training error does not decrease fast enough, start again
  – Repeat multiple times



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

What Properties Facilitate Optimization?

- What properties of the network architecture facilitate optimization?
  - Positive homogeneity
  - Parallel subnetwork structure
- What properties of the regularization function facilitate optimization?
  - Positive homogeneity
  - Adapt network structure to the data [1]



Picture courtesy of Ben Haeffele

[1] Bengio, et al., "Convex neural networks." NIPS. (2005)

Main Results

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

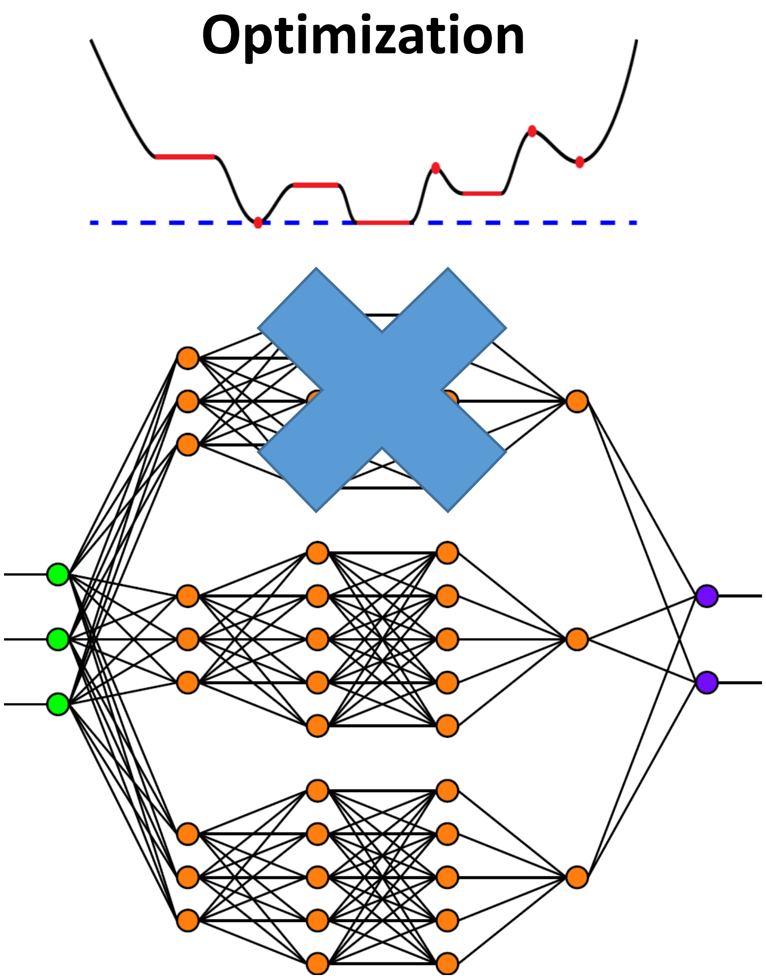
- **Assumptions:**
  - $\ell(Y,X)$: convex and once differentiable in X
  - Φ and Θ: sums of positively homogeneous functions of same degree

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

- **Examples:**
  - ReLU: $\max(\alpha x, 0) = \alpha \max(x,0) \quad \alpha \geq 0$
  - Max pooling: $\max(\alpha x_1,\ldots,\alpha x_D) = \alpha \max(x_1,\ldots,x_D) \quad \alpha \geq 0$
  - Matrix product: $\phi(X^1,X^2) = X^1 {X^2}^\top$
  - Tensor product: $\phi(X^1,\ldots,X^K) = X^1 \otimes \cdots \otimes X^K$
  - Deep neural network: $\phi(X^1,\ldots,X^K) = \psi_K(\cdots\psi_2(\psi_1(VX^1)X^2)\cdots X^K)$

[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
[3] Haeffele, Vidal. Global optimality in neural network training. CVPR 2017.

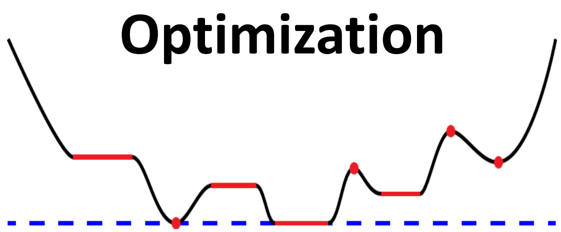
Main Results



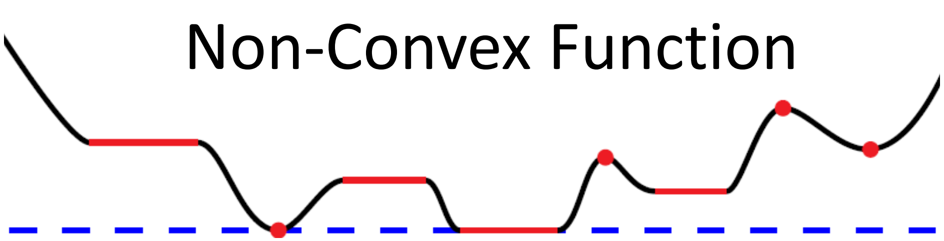
Theorem 1:
A local minimum such that all the weights from one subnetwork are zero is a global minimum

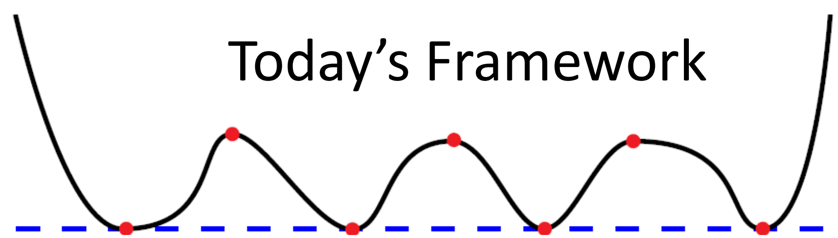
[1] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14
[2] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv, '15
[3] Haeffele, Vidal. Global optimality in neural network training. CVPR 2017.

Main Results



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

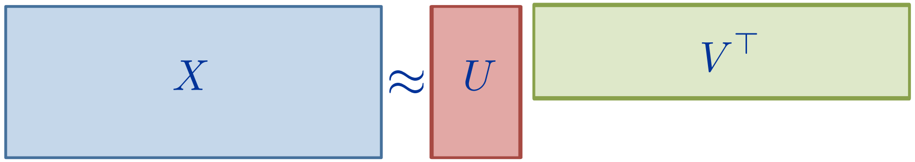




[1] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14
[2] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv, '15
[3] Haeffele, Vidal. Global optimality in neural network training. CVPR 2017.

Outline

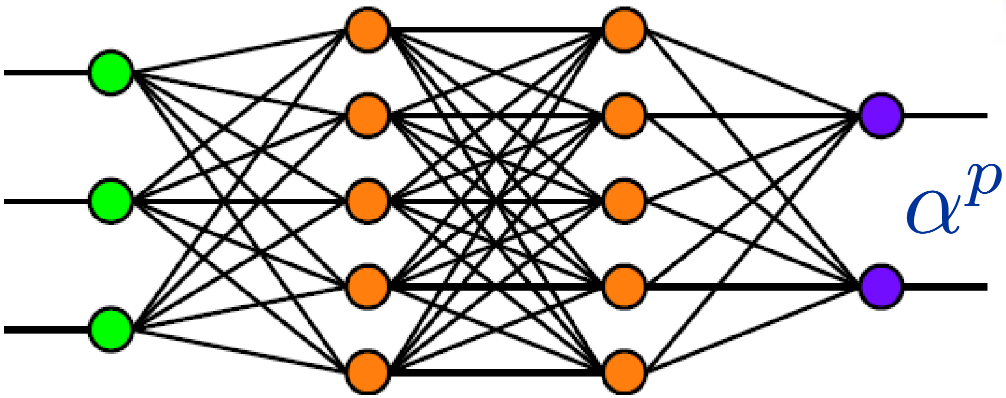
- **Architecture properties that facilitate optimization**
  - Positive homogeneity
  - Parallel subnetwork structure
- **Regularization properties that facilitate optimization**
  - Positive homogeneity
  - Adapt network structure to the data
- **Theoretical guarantees**
  - Sufficient conditions for global optimality
  - Local descent can reach global minimizers



[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
[3] Haeffele, Vidal. Global optimality in neural network training. CVPR 2017.

Key Property #1: Positive Homogeneity

- Start with a network



- Scale the weights by

$$\alpha \geq 0$$

$$\alpha X^1 \quad \alpha X^2 \quad \alpha X^3$$

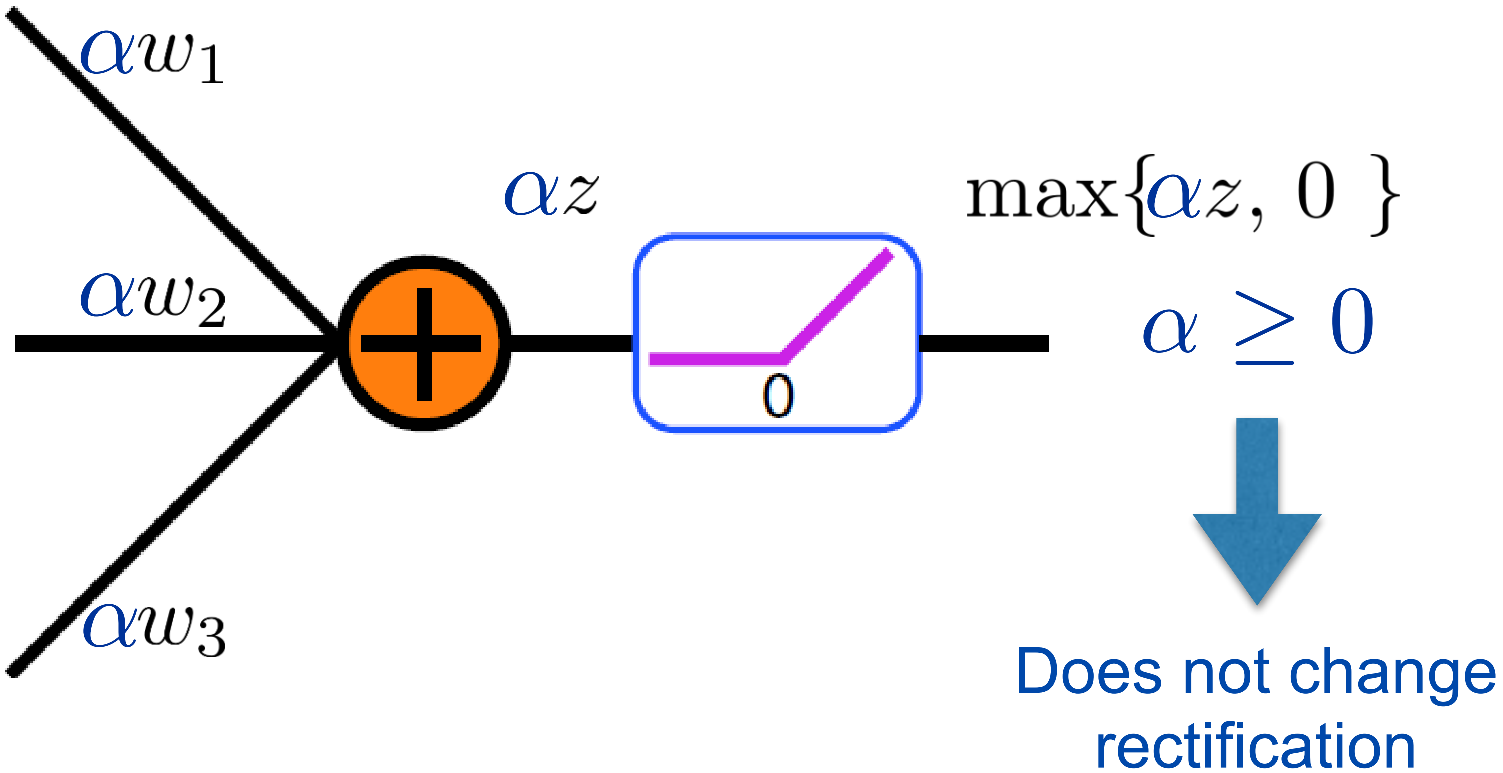
- Output is scaled by α^p, where p = degree of homogeneity

$$\Phi(X^1, X^2, X^3) = Y$$

$$\Phi(\alpha X^1, \alpha X^2, \alpha X^3) = \alpha^p Y$$

Examples of Positively Homogeneous Maps

- **Example 1**: Rectified Linear Units (ReLU)

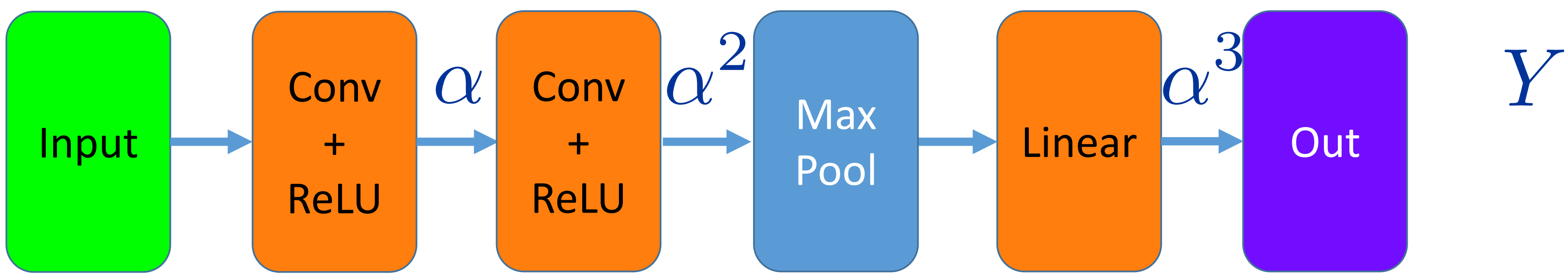


- Linear + ReLU layer is positively homogeneous of degree 1

Examples of Positively Homogeneous Maps

- **Example 2**: Simple networks with convolutional layers, ReLU, max pooling and fully connected layers

$$\max\{\alpha^2 z_1, \alpha^2 z_2\}$$

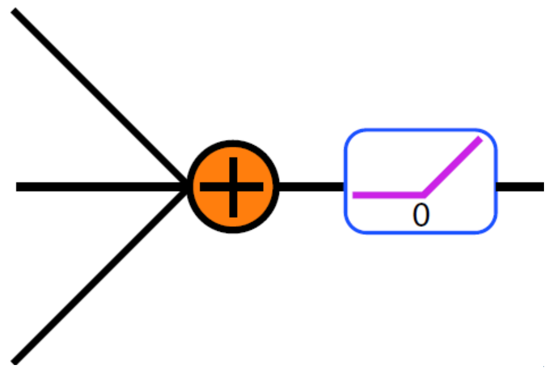


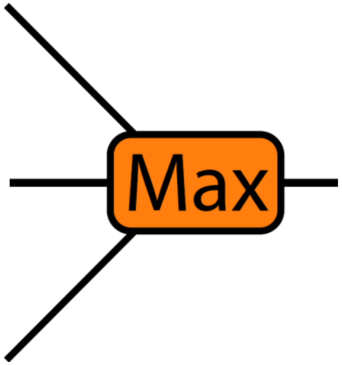
$$\alpha X^1 \quad \alpha X^2 \quad \alpha X^3$$

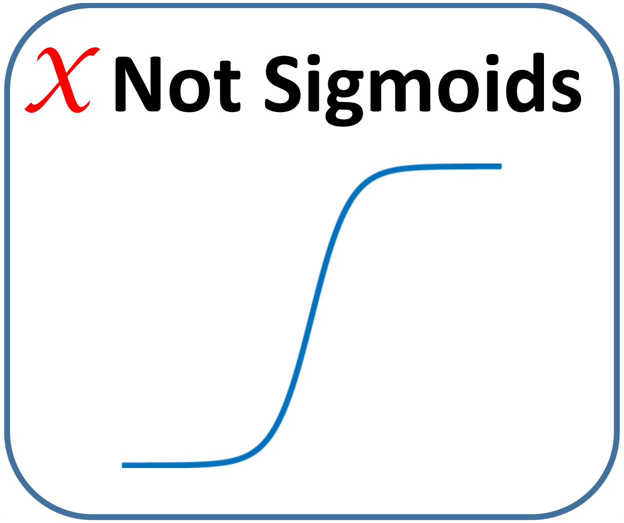
- Typically each weight layer increases degree of homogeneity by 1

Examples of Positively Homogeneous Maps

- Some Common Positively Homogeneous Layers
  - Fully Connected + ReLU
  - Convolution + ReLU
  - Max Pooling
  - Linear Layers
  - Mean Pooling
  - Max Out
  - Many possibilities...





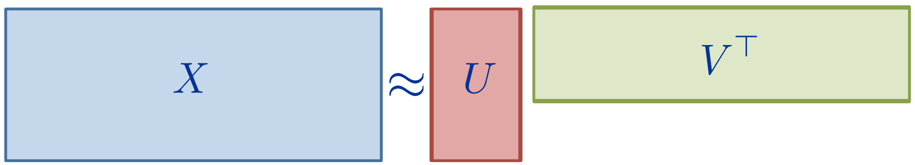


Outline

- **Architecture properties that facilitate optimization**
  - Positive homogeneity
  - Parallel subnetwork structure

- **Regularization properties that facilitate optimization**
  - Positive homogeneity
  - Adapt network structure to the data

- **Theoretical guarantees**
  - Sufficient conditions for global optimality
  - Local descent can reach global minimizers

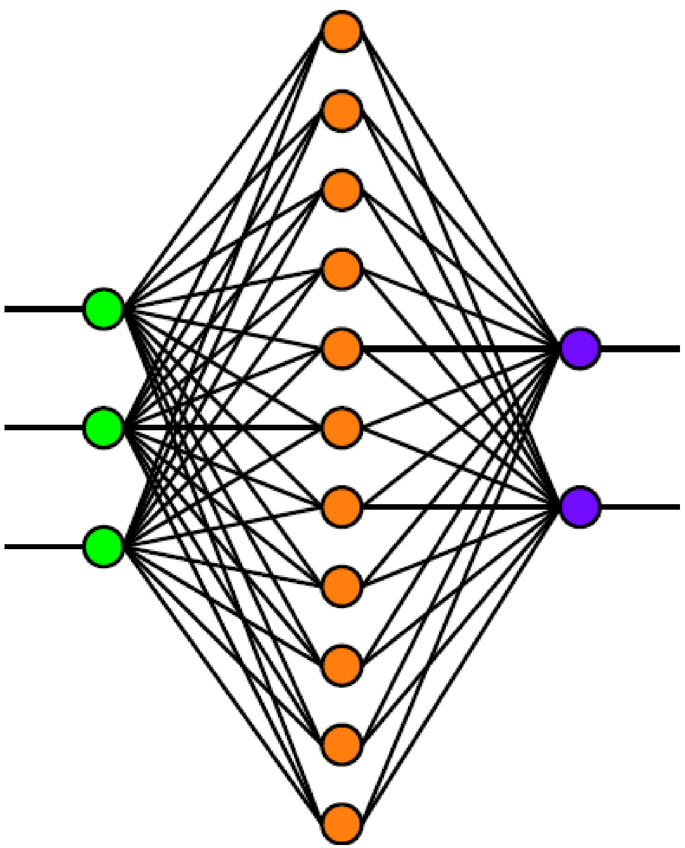


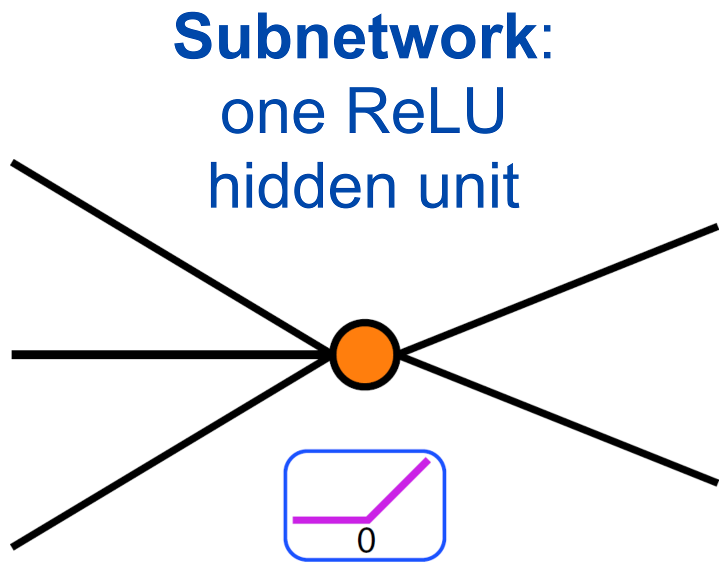
[1] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML ’14
[2] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv, ’15
[3] Haeffele, Vidal. Global optimality in neural network training. CVPR 2017.

Key Property #2: Parallel Subnetworks

- Subnetworks with identical structure connected in parallel

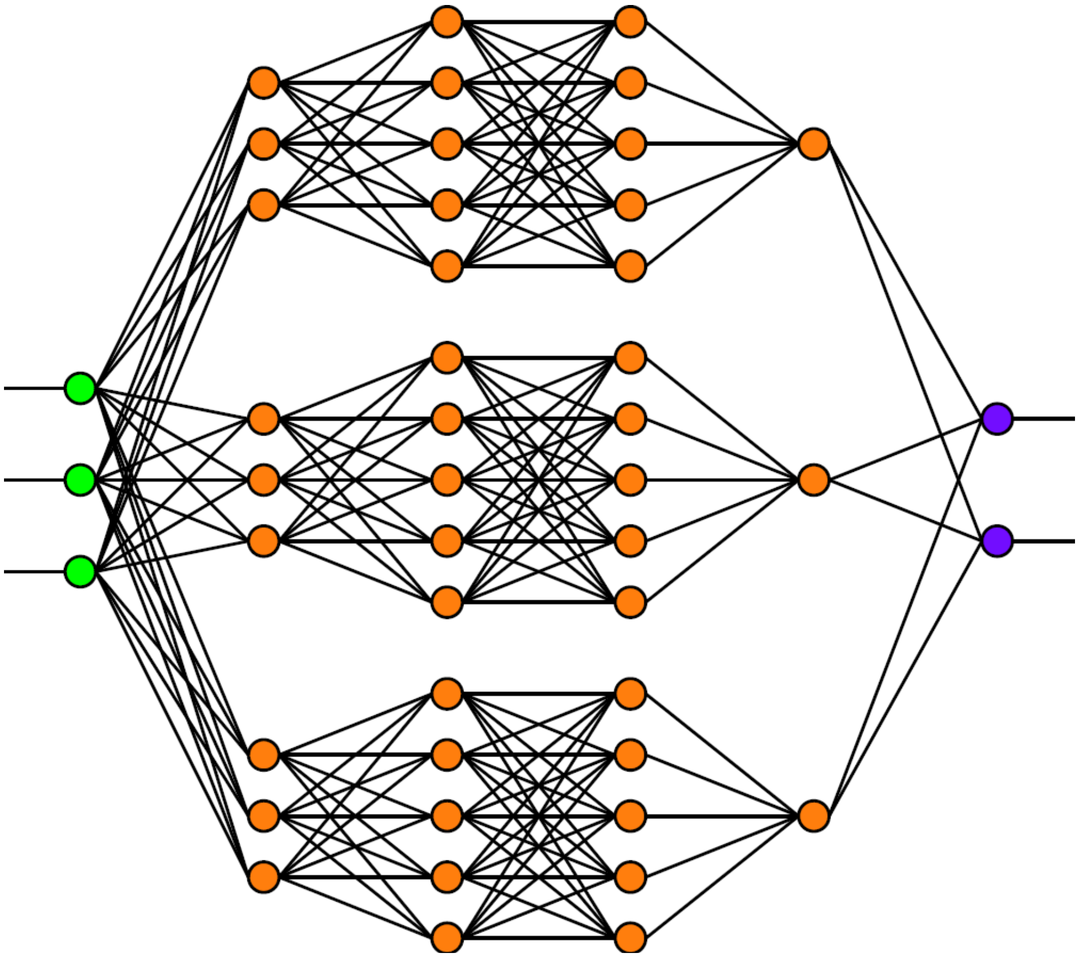
- **Simple example**: single hidden network



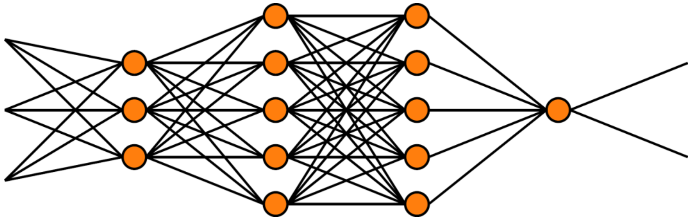


Key Property #2: Parallel Subnetworks

- Any positively homogeneous network can be used

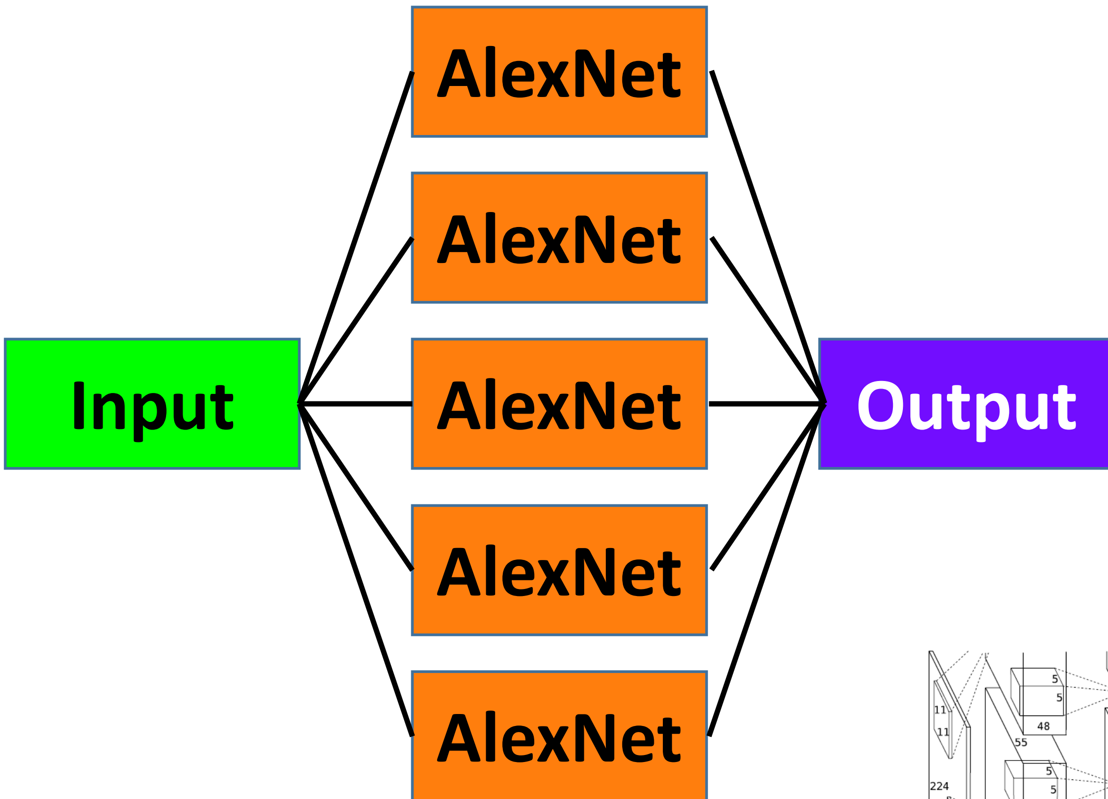


Subnetwork:
multiple
ReLU layers

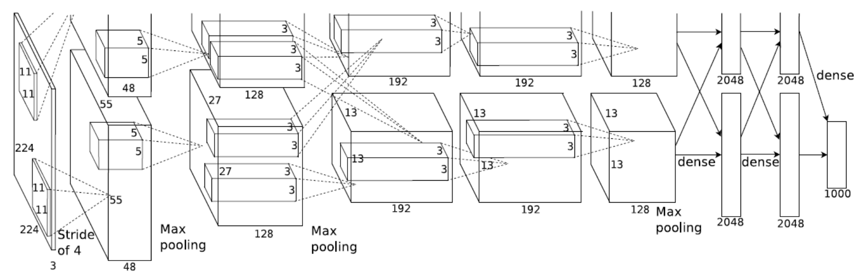


Key Property #2: Parallel Subnetworks

- **Example**: Parallel AlexNets [1]



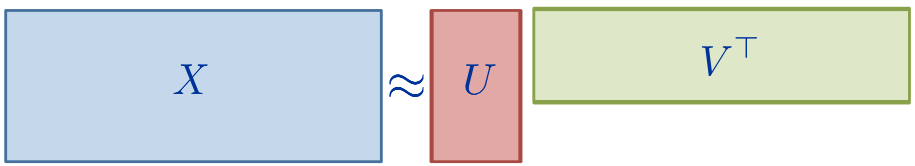
Subnetwork:
AlexNet



[1] Krizhevsky, Sutskever, and Hinton. "Imagenet classification with deep convolutional neural networks." NIPS, 2012

Outline

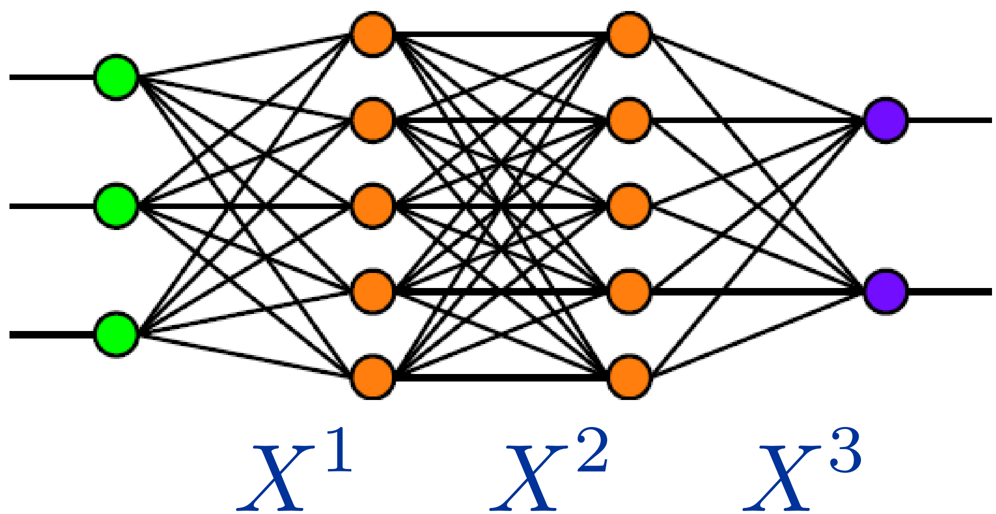
- **Architecture properties that facilitate optimization**
  - Positive homogeneity
  - Parallel subnetwork structure
- **Regularization properties that facilitate optimization**
  - Positive homogeneity
  - Adapt network structure to the data
- **Theoretical guarantees**
  - Sufficient conditions for global optimality
  - Local descent can reach global minimizers



[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
[3] Haeffele, Vidal. Global optimality in neural network training. CVPR 2017.

Basic Regularization: Weight Decay

$$\Theta(X^1, X^2, X^3) = \|X^1\|_F^2 + \|X^2\|_F^2 + \|X^3\|_F^2$$

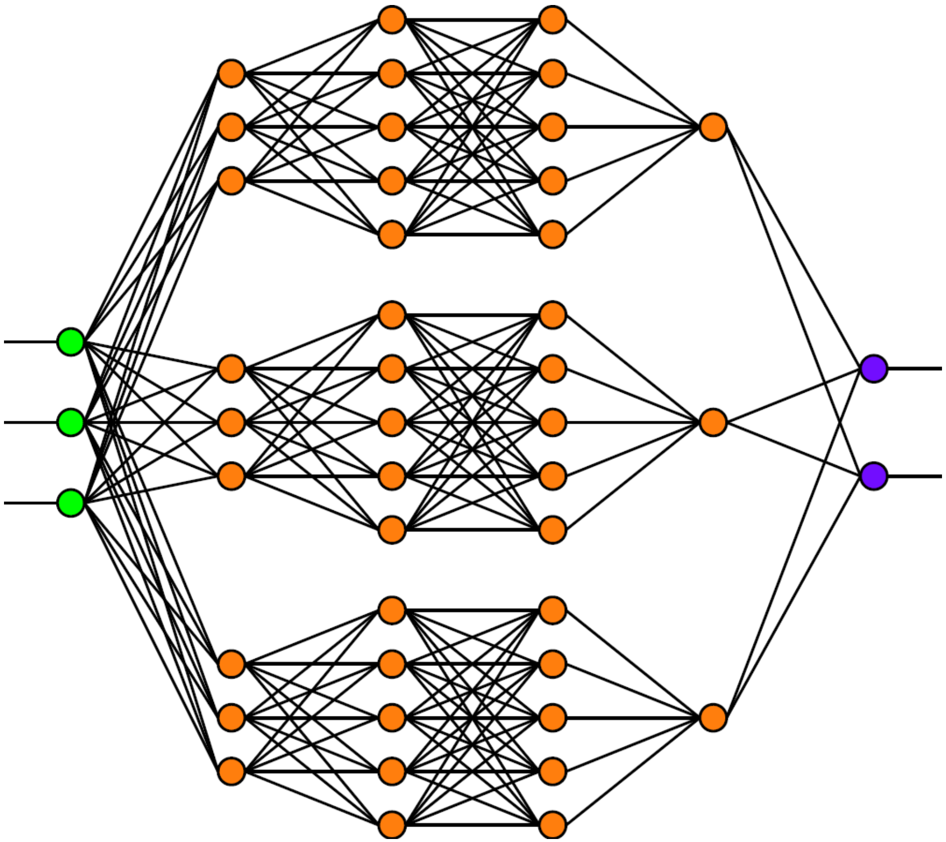


$$\Theta(\alpha X^1, \alpha X^2, \alpha X^3) = \alpha^2 \Theta(X^1, X^2, X^3)$$
$$\Phi(\alpha X^1, \alpha X^2, \alpha X^3) = \alpha^3 \Phi(X^1, X^2, X^3)$$

- **Proposition** non-matching degrees => spurious local minima

Regularizer Adapted to Network Size

- Start with a positively homogeneous network with parallel structure

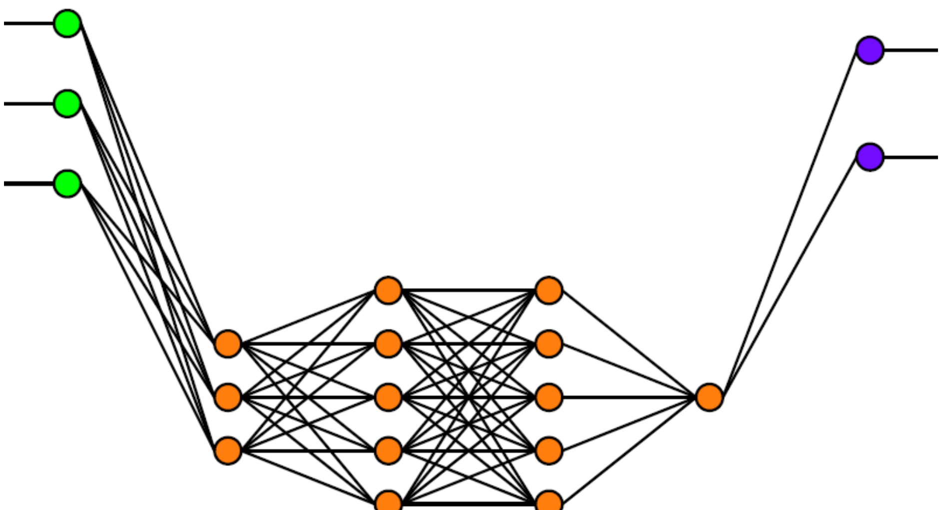


Regularizer Adapted to Network Size

- Take weights of one subnetwork

- Define a regularizer

$$\theta(X_1^1, X_1^2, X_1^3, X_1^4, X_1^5)$$



$X_1^1 \quad X_1^2 \quad X_1^3 \quad X_1^4 \quad X_1^5$

- Nonnegative
- Positively homogeneous with the same degree as network

$$\Phi(\alpha X) = \alpha^p \Phi(X)$$

$$\theta(\alpha X) = \alpha^p \theta(X)$$

- **Example:** product of norms

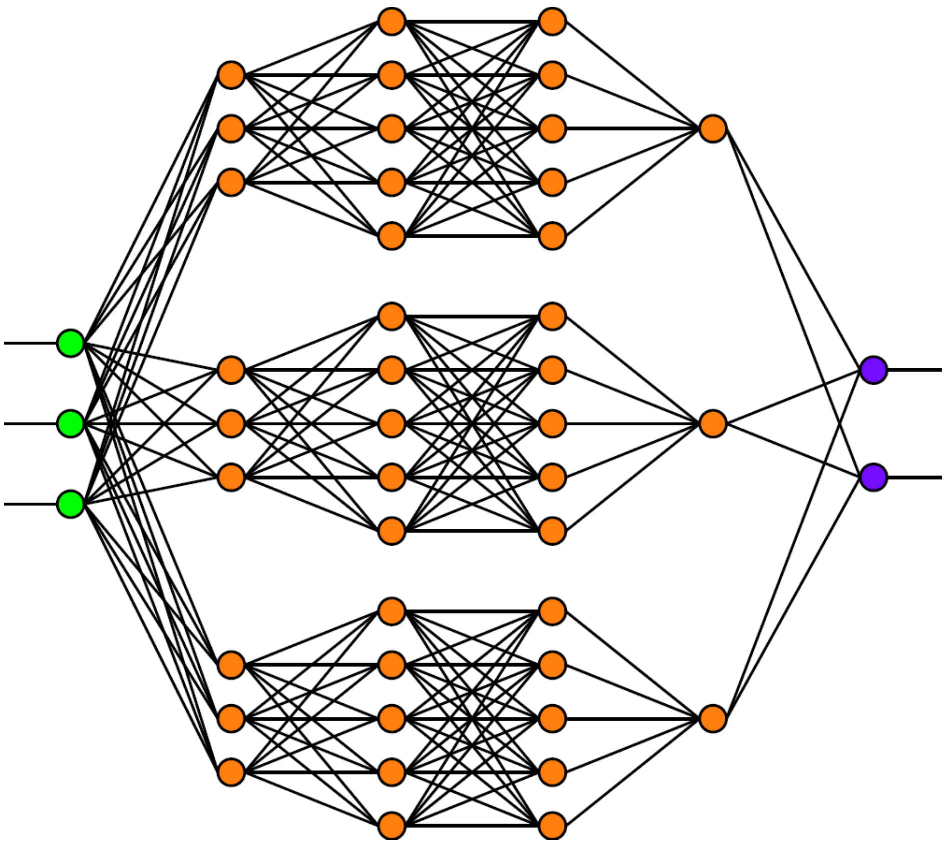
$$\|X_1^1\| \|X_1^2\| \|X_1^3\| \|X_1^4\| \|X_1^5\|$$

Regularizer Adapted to Network Size

- Sum over all subnetworks

$$\Theta(X) = \sum_{i=1}^{r} \theta(X^i)$$

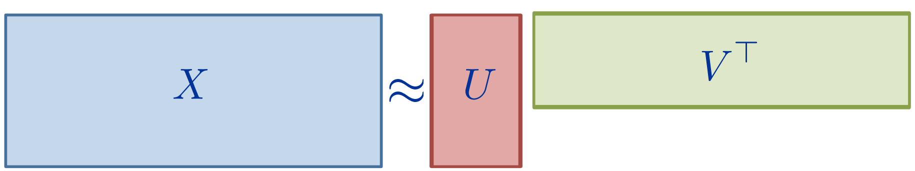
$$r = \#\ \text{subnets}$$



- Allow r to vary
- Adding a subnetwork is penalized by an additional term in the sum
- Regularizer constraints number of subnetworks

Outline

- **Architecture properties that facilitate optimization**
  - Positive homogeneity
  - Parallel subnetwork structure
- **Regularization properties that facilitate optimization**
  - Positive homogeneity
  - Adapt network structure to the data
- **Theoretical guarantees**
  - Sufficient conditions for global optimality
  - Local descent can reach global minimizers



[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
[3] Haeffele, Vidal. Global optimality in neural network training. CVPR 2017.

JHU vision lab

Global Optimality in Structured Matrix Factorization

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

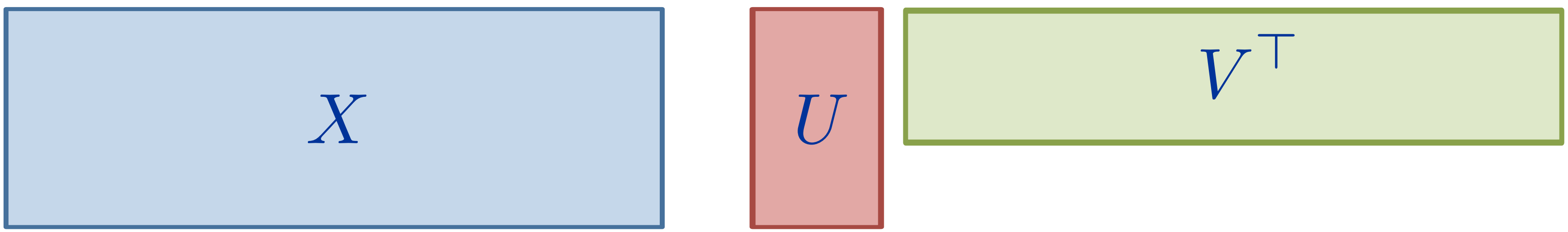
Typical Low-Rank Formulations

- **Convex formulations**:

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

- **Factorized formulations**:

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



Convex formulations:

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

✓ Convex
✳ Large problem size
✳ Unstructured factors

Factorized formulations:

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

✳ Non-Convex
✓ Small problem size
✓ Structured factors

Relating Convex & Factorized Formulations

- **Convex formulations:** **Factorized formulations**

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

- Variational form of the nuclear norm [1,2]

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

$$\boxed{\|X\|_* = \sum \sigma_i(X)}$$

- A natural generalization is the projective tensor norm [3,4]

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

[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.
[3] Bach, Mairal, Ponce, Convex sparse matrix factorizations, arXiv 2008.
[4] Bach. Convex relaxations of structured matrix factorizations, arXiv 2013.

Main Results: Projective Tensor Norm Case

- **Theorem 1**: Assume ℓ is convex and once differentiable in X. A local minimizer (U, V) of the non-convex 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. Moreover, $UV^\top$ is a global minimizer of the convex problem

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

- **Proof sketch**:
  - Convex problem gives global lower bound for non-convex problem
  - If (U, V) local min. of non-convex, then $UV^\top$ global min. of convex

[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

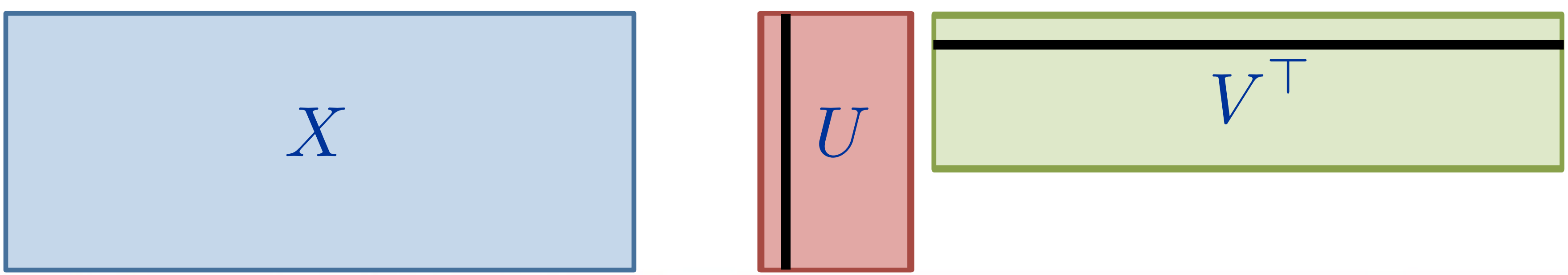
Main Results: Projective Tensor Norm Case

- **Theorem 1**: Assume ℓ is convex and once differentiable in X. A local minimizer (U,V) of the non-convex 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. Moreover, $UV^\top$ is a global minimizer of the convex problem

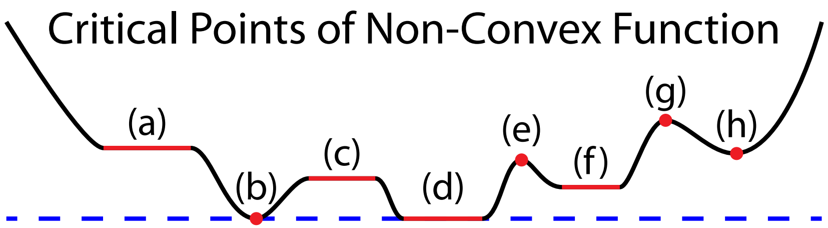
$$\min_{X} \ell(Y, X) + \lambda \|X\|_{u,v}$$

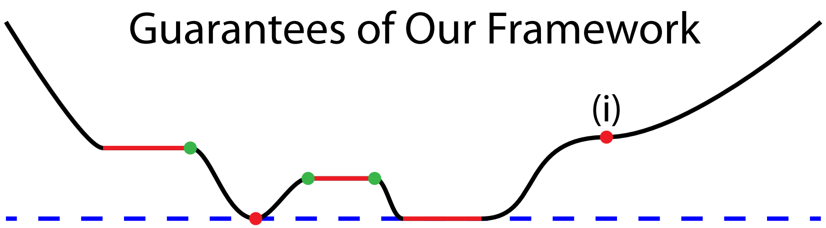


[1] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14
[2] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv '15

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

Main Results: Homogeneous Regularizers

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

- **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

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)

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

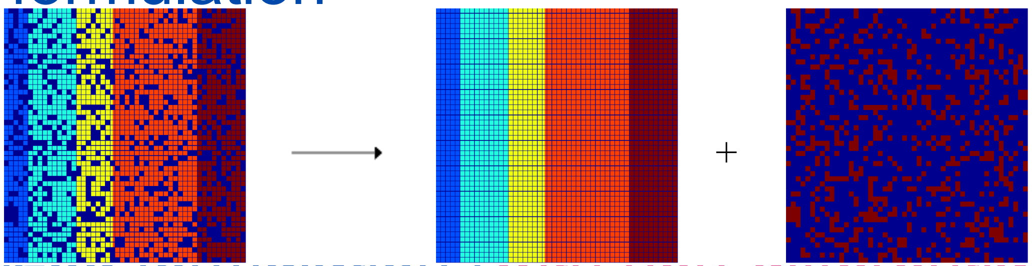
$$\min_{U,V} \|Y - UV^\top\|_F^2 + \lambda \sum_i |U_i|_2(|V_i|_2 + \gamma|V_i|_1)$$

Example: Robust PCA

- Original formulation [1]

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

- Equivalent formulation



- New factorized formulation (with non-differentiable loss)

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

- New factorized formulation (with differentiable loss)

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

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

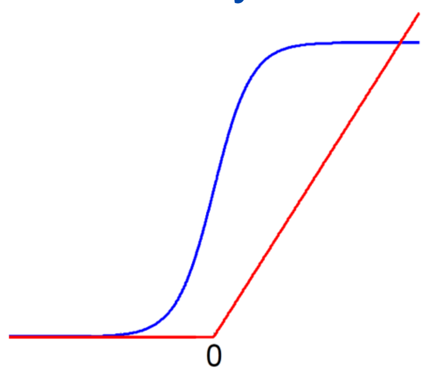
JHU vision lab

Global Optimality in Positively Homogeneous Factorization

René Vidal
Center for Imaging Science
Institute for Computational Medicine

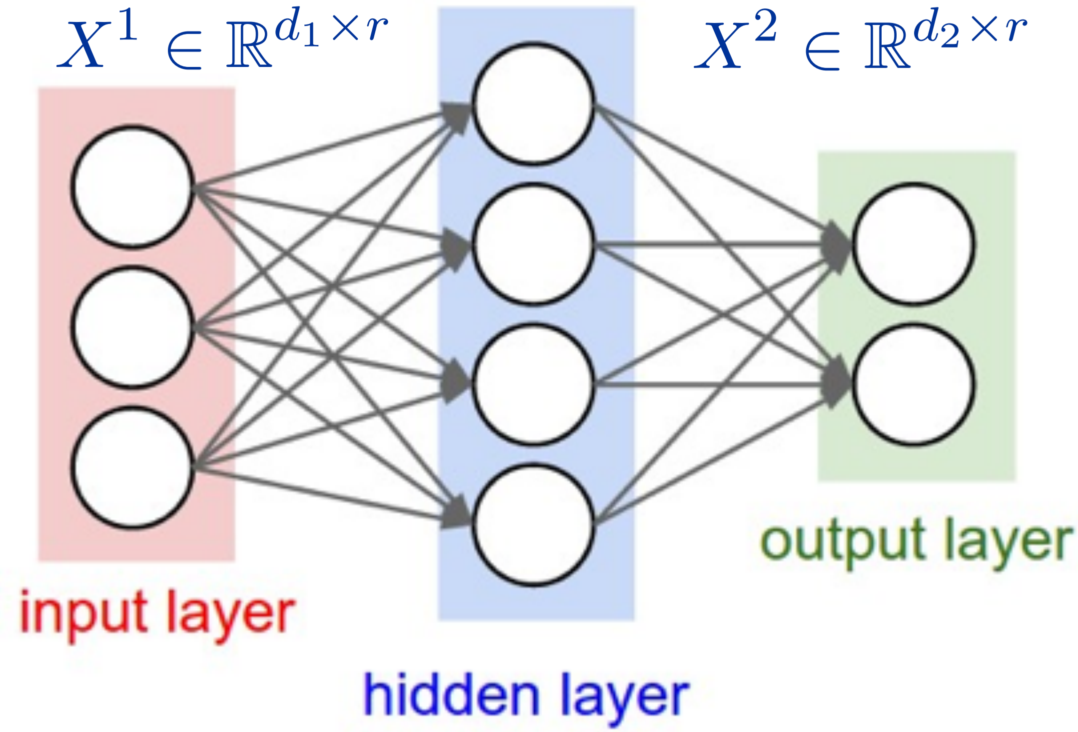
From Matrix Factorizations to Deep Learning

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



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

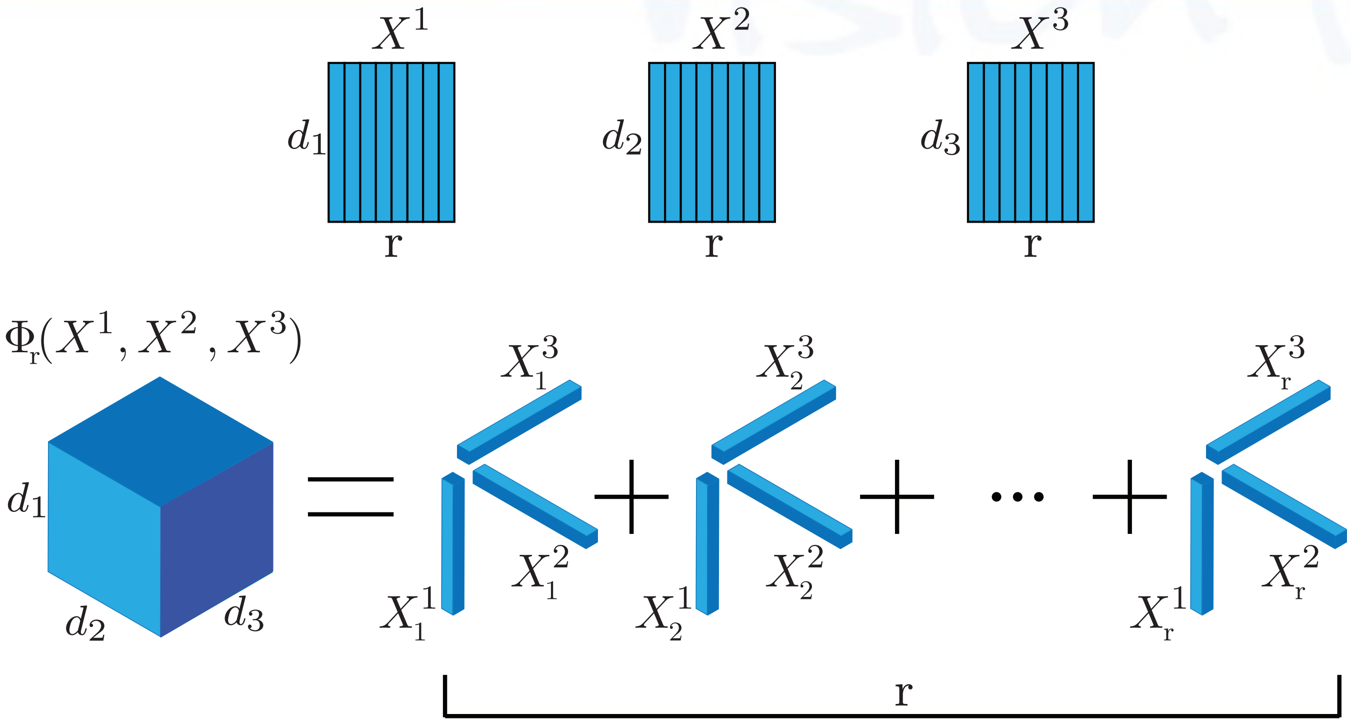
$X^1 \in \mathbb{R}^{d_1 \times r}$ $X^2 \in \mathbb{R}^{d_2 \times r}$



- "Almost" like matrix factorization
  - r = rank
  - r = #neurons in hidden layer
  - ReLU + max pooling is positively homogeneous of degree 1

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

From Matrix to Tensor Factorization



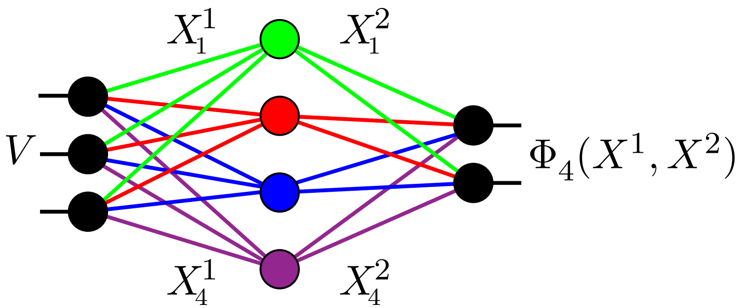
- Tensor product $\phi(X^1, \ldots, X^K) = X^1 \otimes \cdots \otimes X^K$ is positively homogeneous of degree K

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

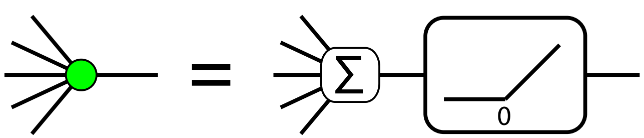
From Matrix Factorizations to Deep Learning

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

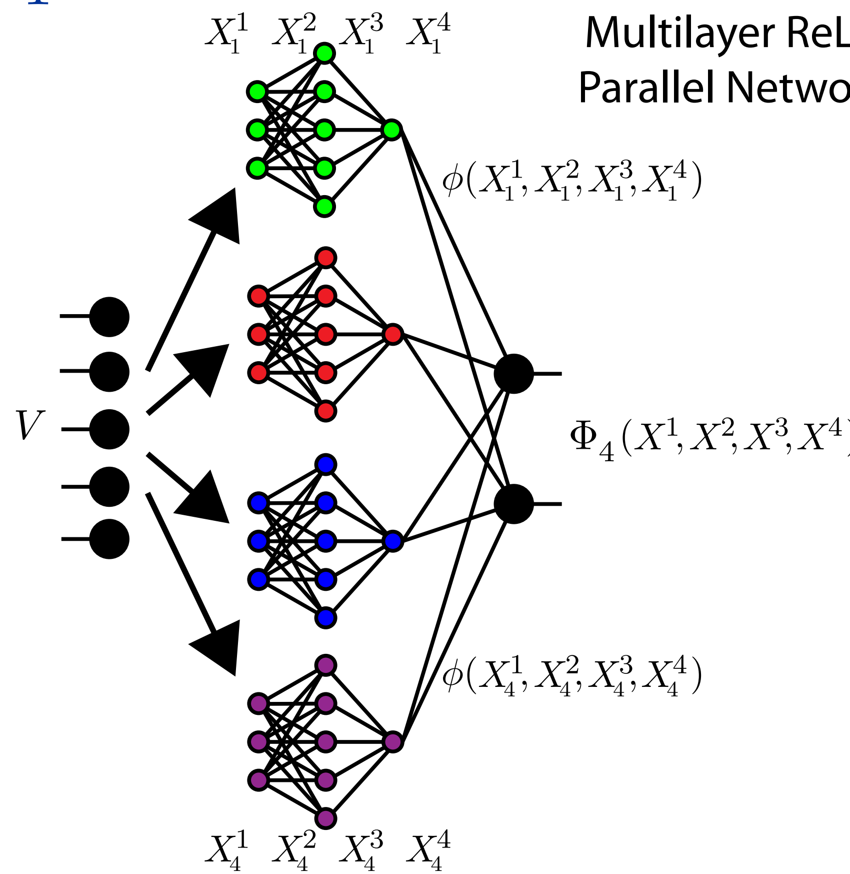
ReLU Network with One Hidden Layer



Rectified Linear Unit (ReLU)



Multilayer ReLU Parallel Network



Key Ingredient: Proper Regularization

- 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"

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

  where θ is positively homogeneous of the same degree as ϕ

- **Proposition:** $\Omega_{\phi,\theta}$ is convex
- **Intuition:** regularizer Θ "comes from a convex function"

Main Results

- **Theorem 1:** Assume ℓ is convex and once differentiable in X. A local minimizer $(X^1, \ldots, X^K)$ of the factorized formulation

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

  such that for some i and all k $X_i^k = 0$ is a global minimizer. Moreover, $X = \Phi(X^1, \ldots, X^K)$ is a global minimizer of the convex problem

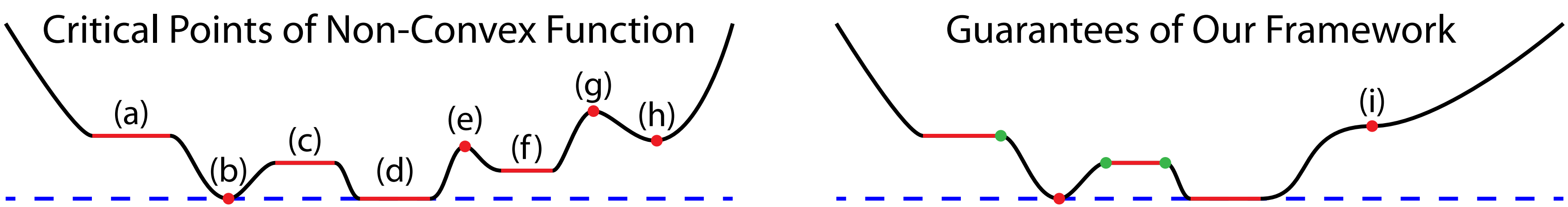
$$\min_X \ell(Y, X) + \lambda \Omega_{\phi,\theta}(X)$$

- 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 parallel) 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

Experimental Results

- Better performance with less training examples [Sokolic, Giryes, Sapiro, Rodrigues, 2017]
  - WD = weight decay
  - LM = Jacobian regularizer ~ product of weights regularizer

		256 samples			512 samples			1024 samples		
loss	# layers	no reg.	WD	LM	no reg.	WD	LM	no reg.	WD	LM
hinge	2	88.37	89.88	**93.83**	93.99	94.62	95.49	95.79	96.57	97.45
hinge	3	87.22	89.31	93.22	93.41	93.97	**95.76**	95.46	96.45	**97.60**
CCE	2	88.45	88.45	92.77	92.29	93.14	95.25	95.38	95.79	96.89
CCE	3	89.05	89.05	93.10	91.81	93.02	95.32	95.11	95.86	97.14

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!