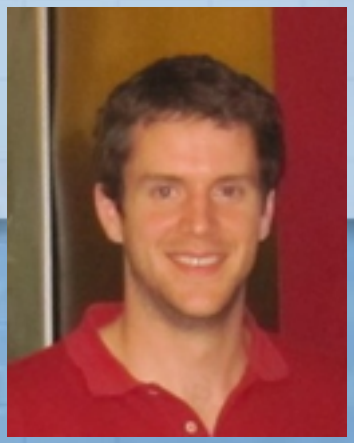
JHU vision lab

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



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

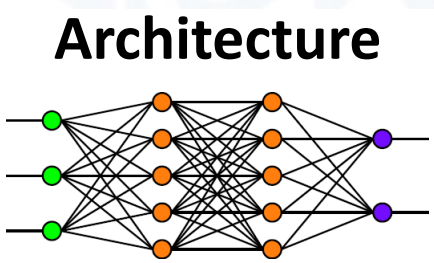
JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

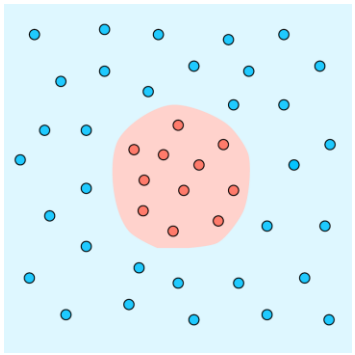
THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins

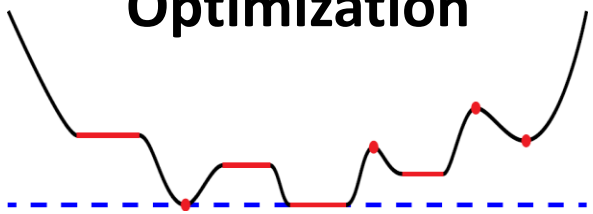
CIS Center for IMAGING SCIENCE

This Talk: Analysis of 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]



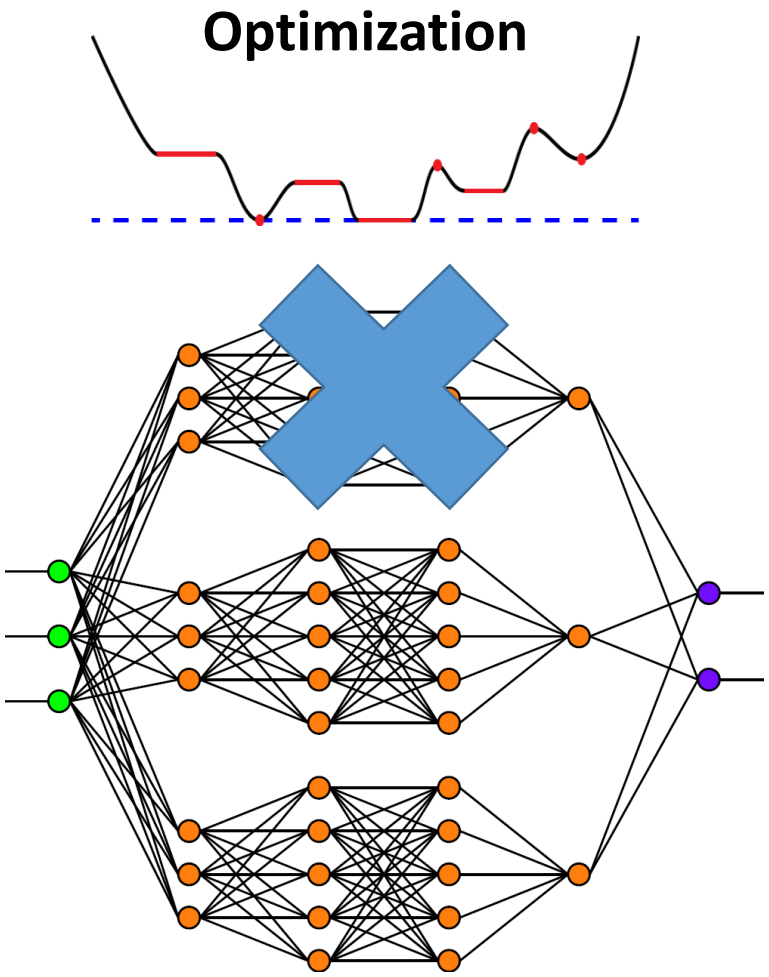




Picture courtesy of Ben Haeffele

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

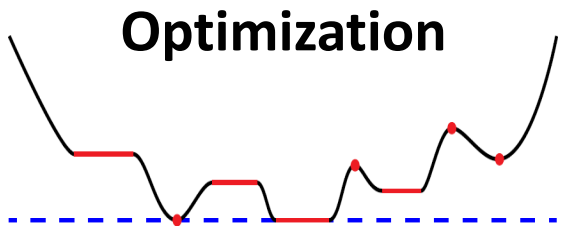
Main Results



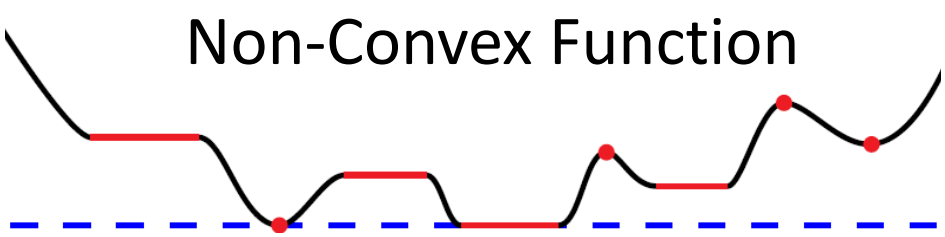
Theorem 1:
A local minimum such that all the weights from one subnetwork are zero is a global minimum

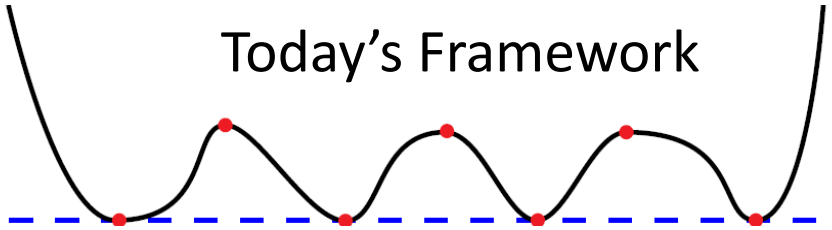
[1] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14
[2] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv, '15
[3] Haeffele, Vidal. Global optimality in neural network training.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

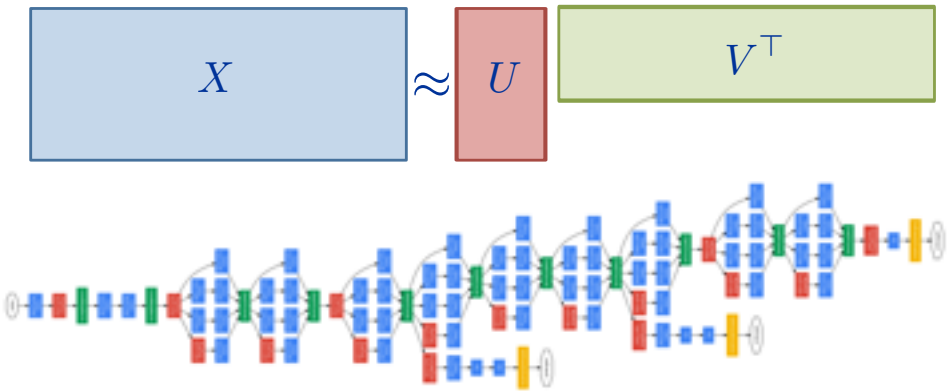




[1] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14
[2] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv, '15
[3] Haeffele, Vidal. Global optimality in neural network training.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

Outline

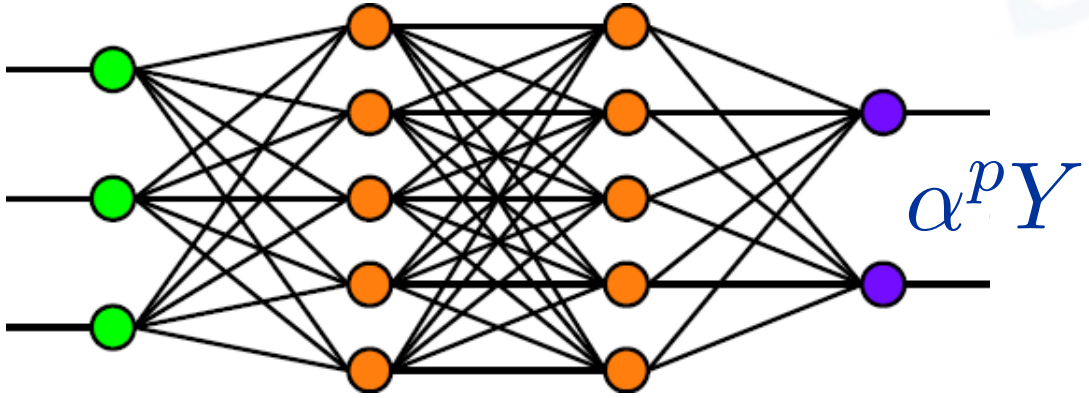
- **Architecture properties that facilitate optimization**
  - Positive homogeneity
  - Parallel subnetwork structure
- **Regularization properties that facilitate optimization**
  - Positive homogeneity
  - Adapt network structure to the data
- **Theoretical guarantees**
  - Sufficient conditions for 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

$$\alpha \geq 0$$

$$\alpha W^1 \;\; \alpha W^2 \;\; \alpha W^3$$

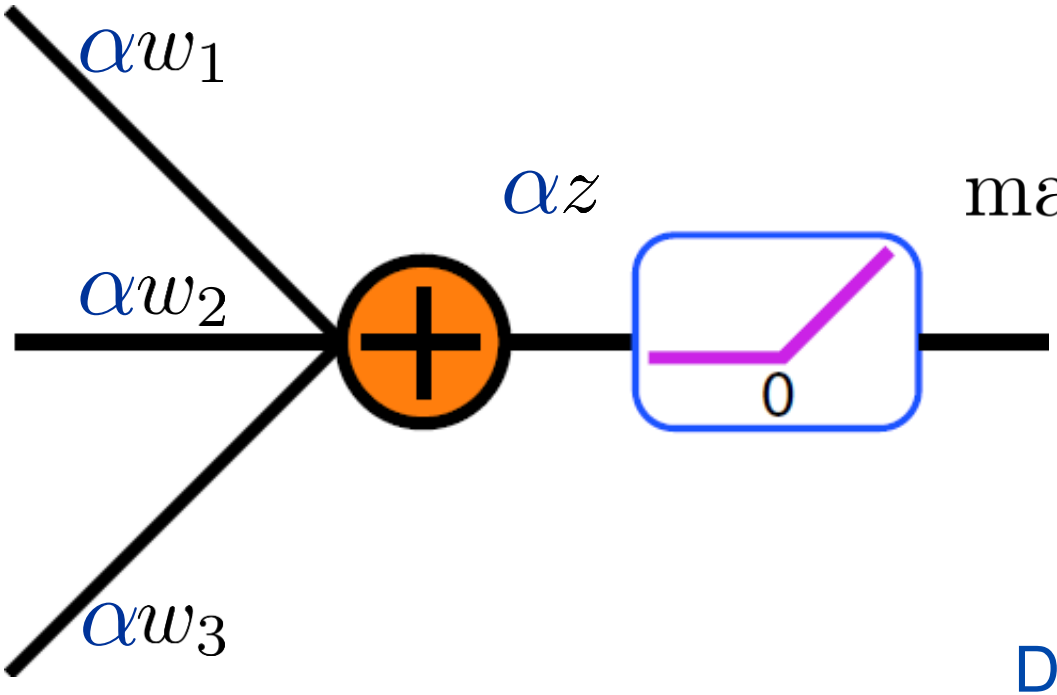
- Output is scaled by α^p, where p = degree of homogeneity

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

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

Examples of Positively Homogeneous Maps

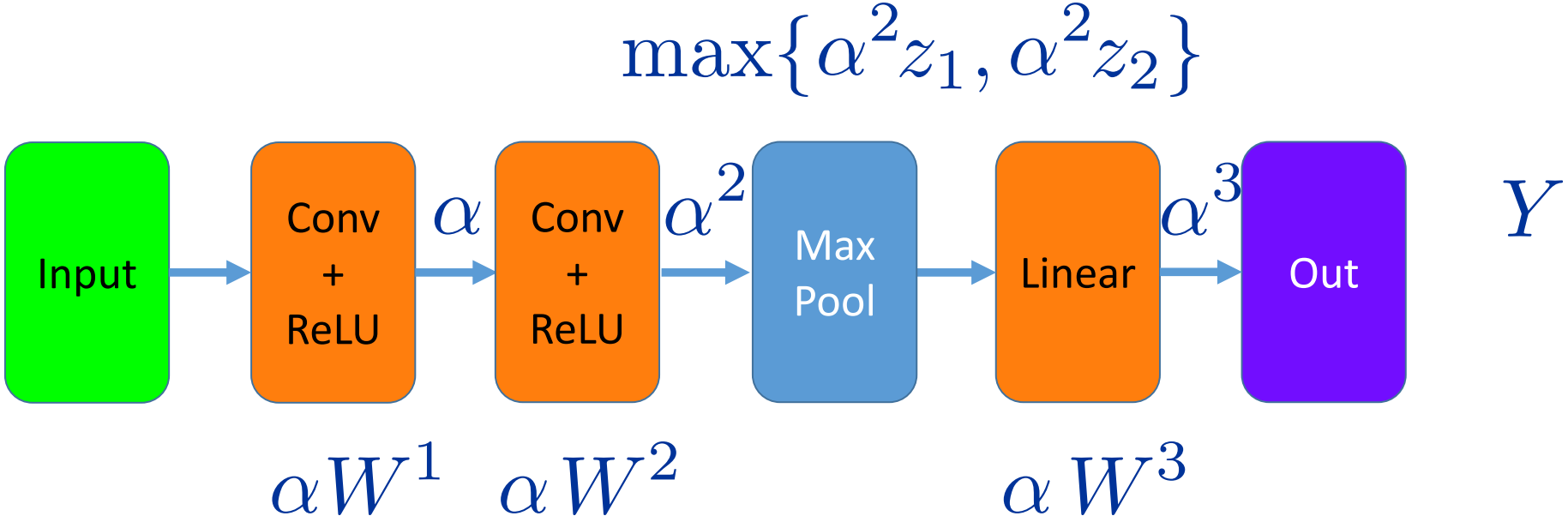
- **Example 1**: Rectified Linear Units (ReLU)



- Linear + ReLU layer is positively homogeneous of degree 1

Examples of Positively Homogeneous Maps

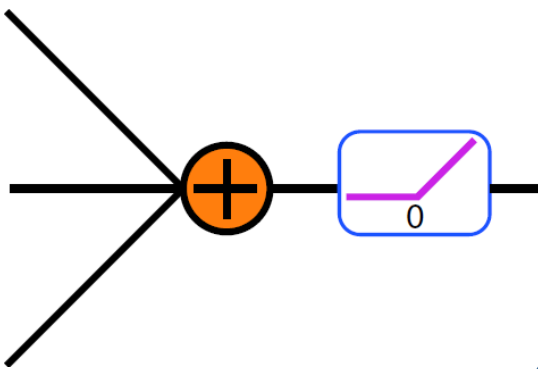
- **Example 2**: Simple networks with convolutional layers, ReLU, max pooling and fully connected layers

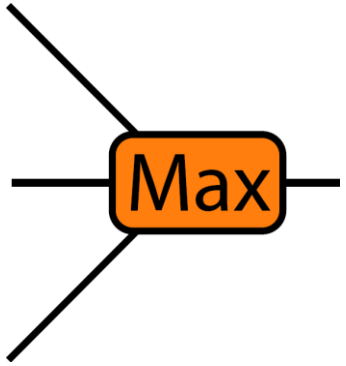


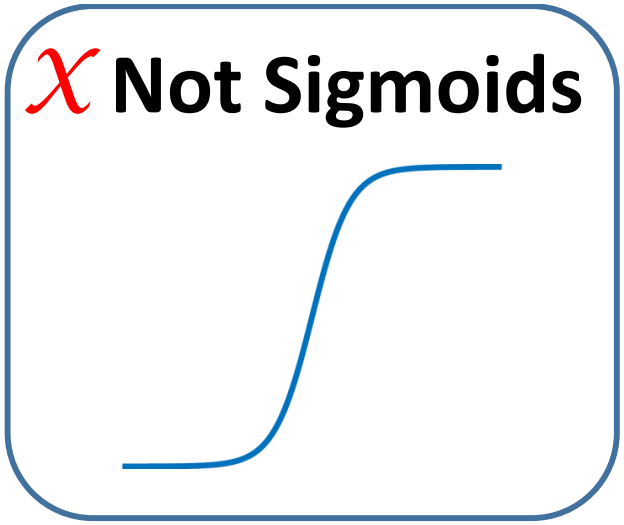
- Typically each weight layer increases degree of homogeneity by 1

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...

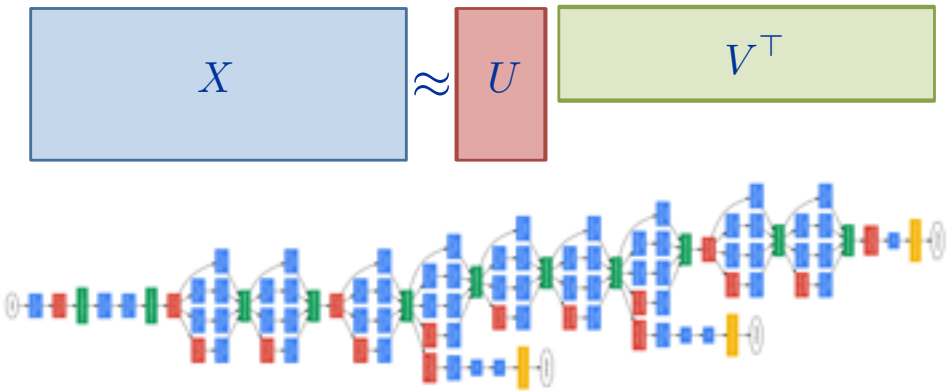






Outline

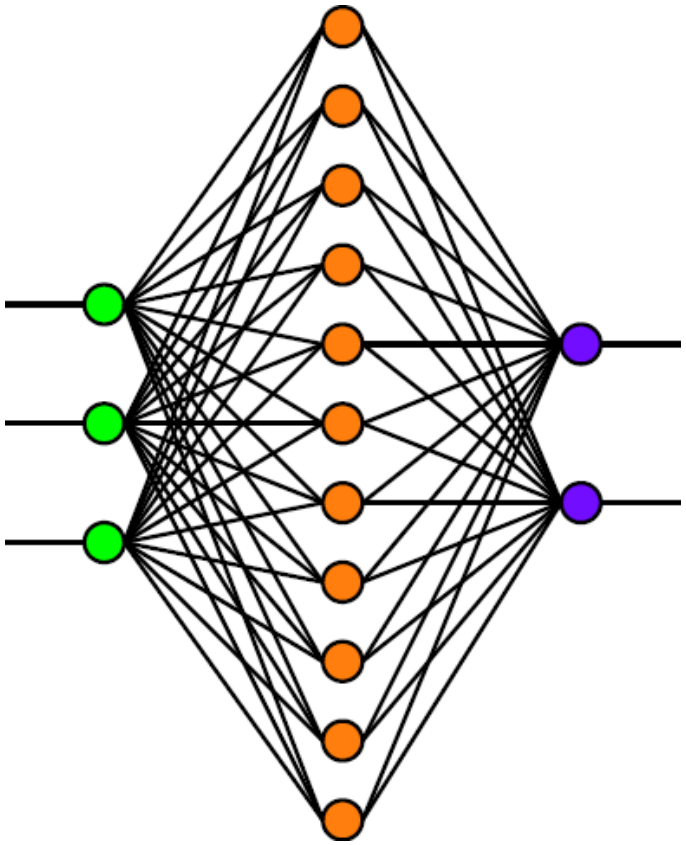
- **Architecture properties that facilitate optimization**
  - Positive homogeneity
  - Parallel subnetwork structure
- **Regularization properties that facilitate optimization**
  - Positive homogeneity
  - Adapt network structure to the data
- **Theoretical guarantees**
  - Sufficient conditions for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



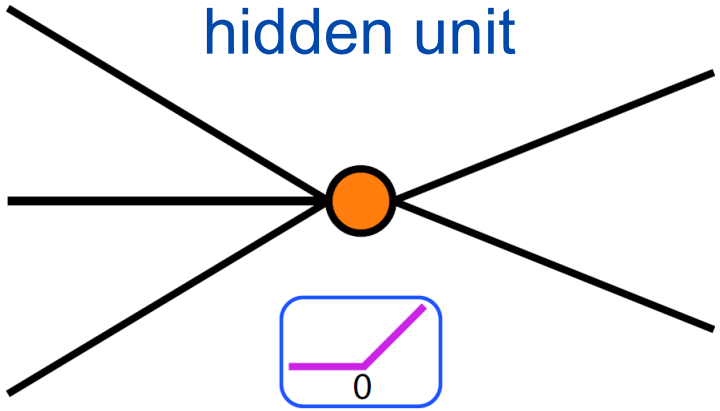
[1] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14
[2] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv, '15
[3] Haeffele, Vidal. Global optimality in neural network training.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

Key Property #2: Parallel Subnetworks

- Subnetworks with identical structure connected in parallel
- **Simple example**: single hidden network

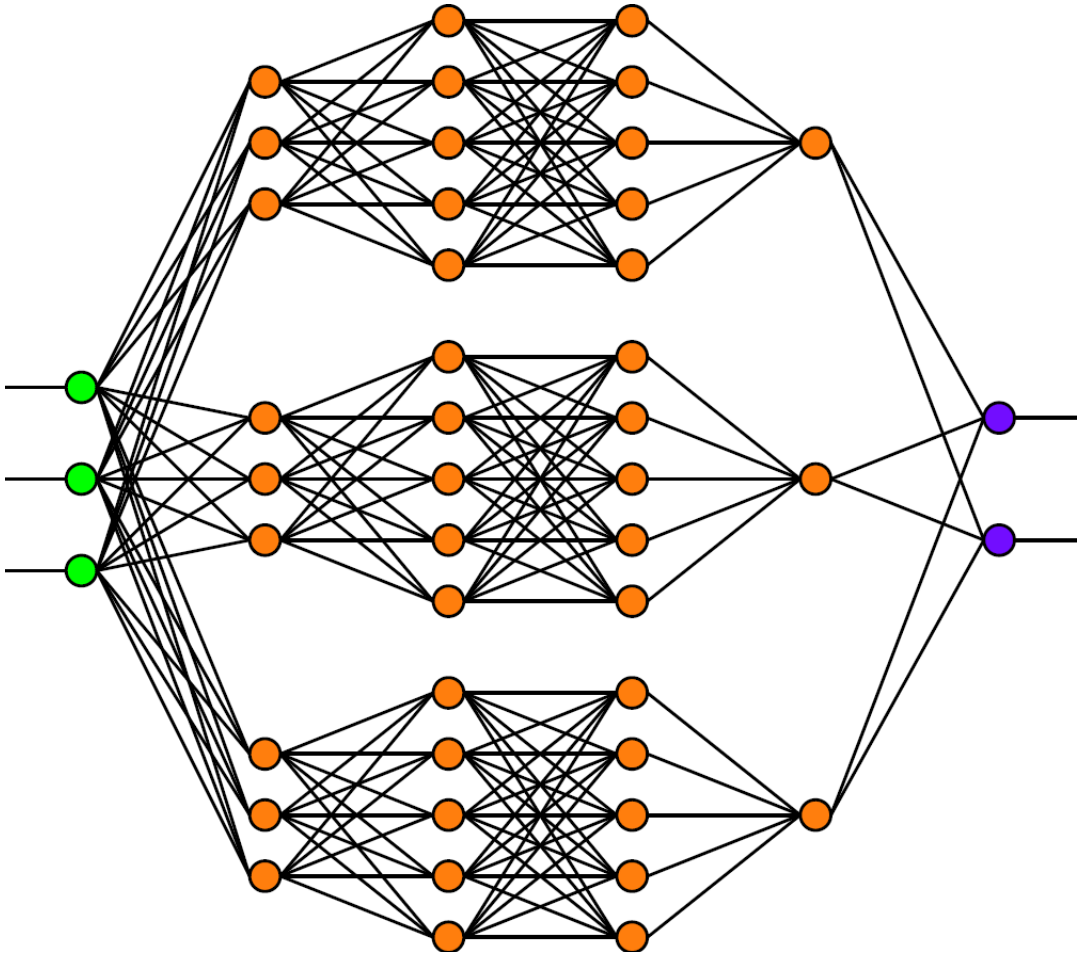


Subnetwork:
one ReLU
hidden unit

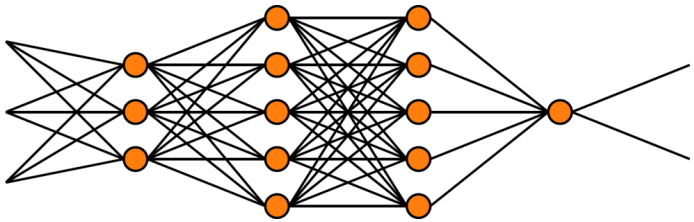


Key Property #2: Parallel Subnetworks

- Any positively homogeneous network can be used

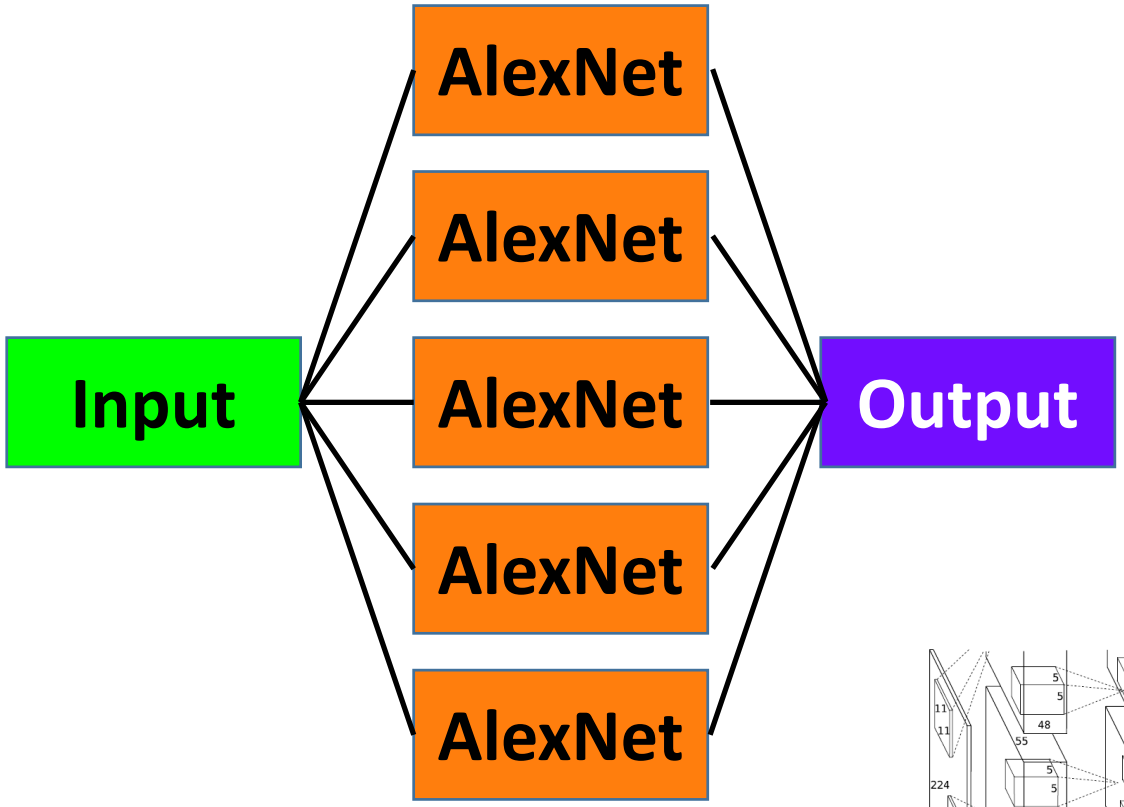


Subnetwork: multiple ReLU layers

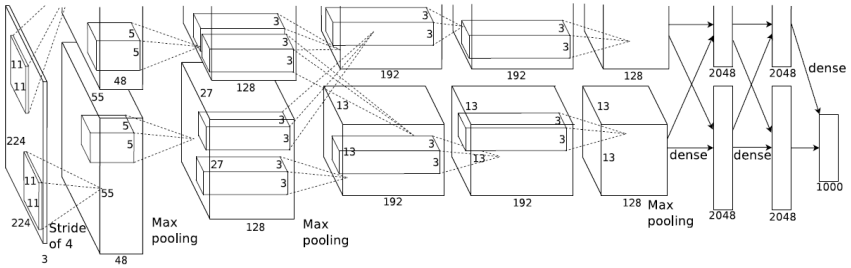


Key Property #2: Parallel Subnetworks

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



Subnetwork:
AlexNet



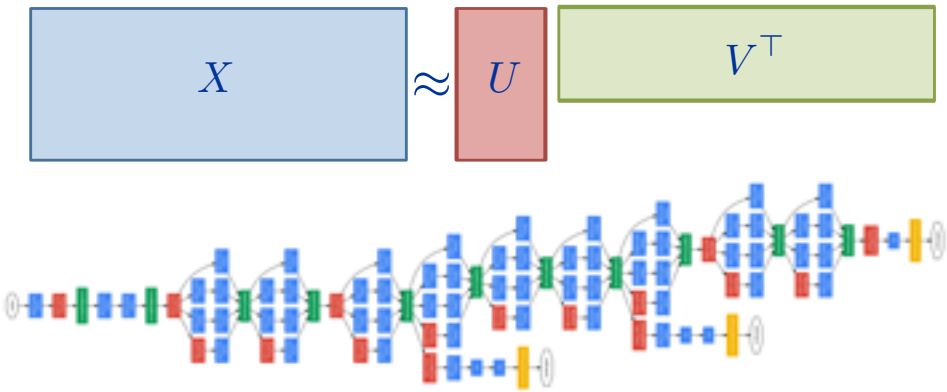
[1] Krizhevsky, Sutskever, and Hinton. "Imagenet classification with deep convolutional neural networks." NIPS, 2012

Outline

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

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

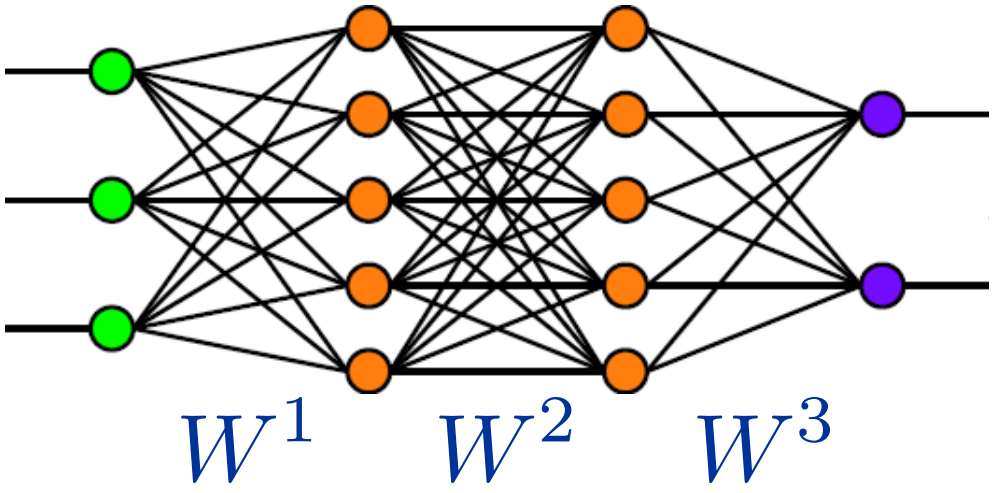
- **Theoretical guarantees**
  - Sufficient conditions for global optimality
  - Local descent can reach global minimizers



[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

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

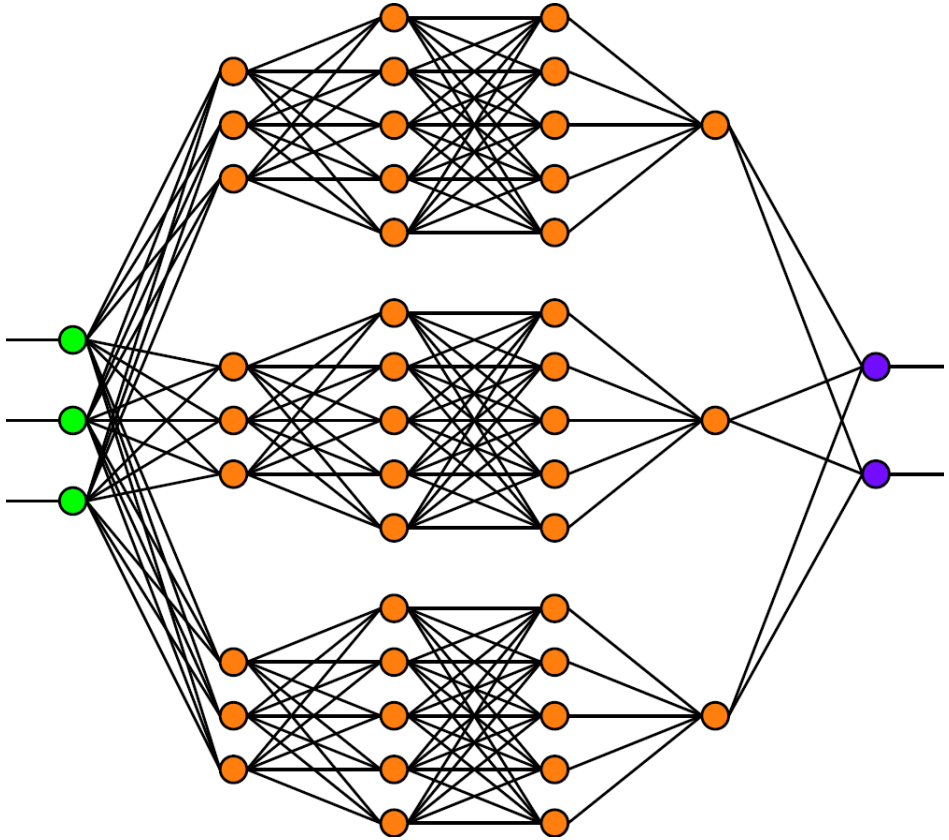


$$\Theta(\alpha W^1, \alpha W^2, \alpha W^3) = \alpha^2 \Theta(W^1, W^2, W^3)$$
$$\Phi(\alpha W^1, \alpha W^2, \alpha W^3) = \alpha^3 \Phi(W^1, W^2, W^3)$$

- **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

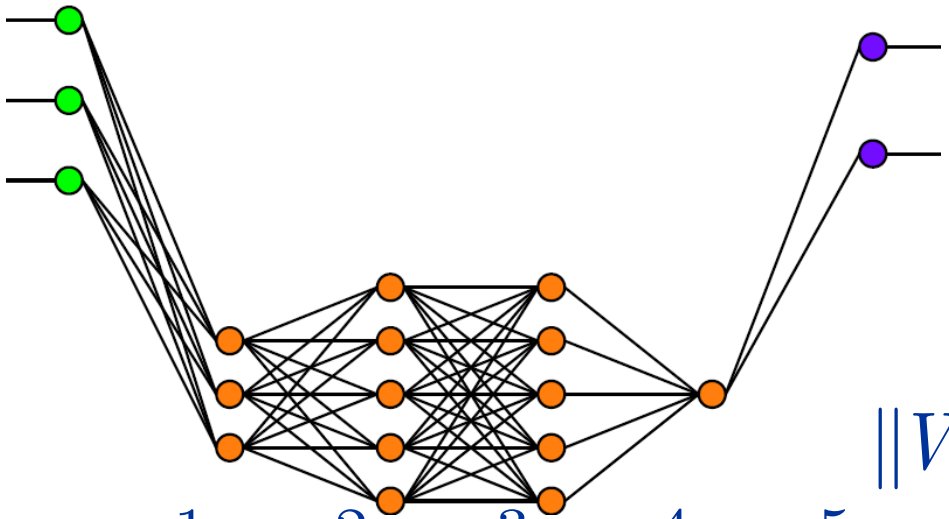
- Take weights of one subnetwork
- Define a regularizer $\theta(W_1^1, W_1^2, W_1^3, W_1^4, W_1^5)$
  - Positive semi-definite
  - Positively homogeneous with the same degree as network

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

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

- **Example:** product of norms

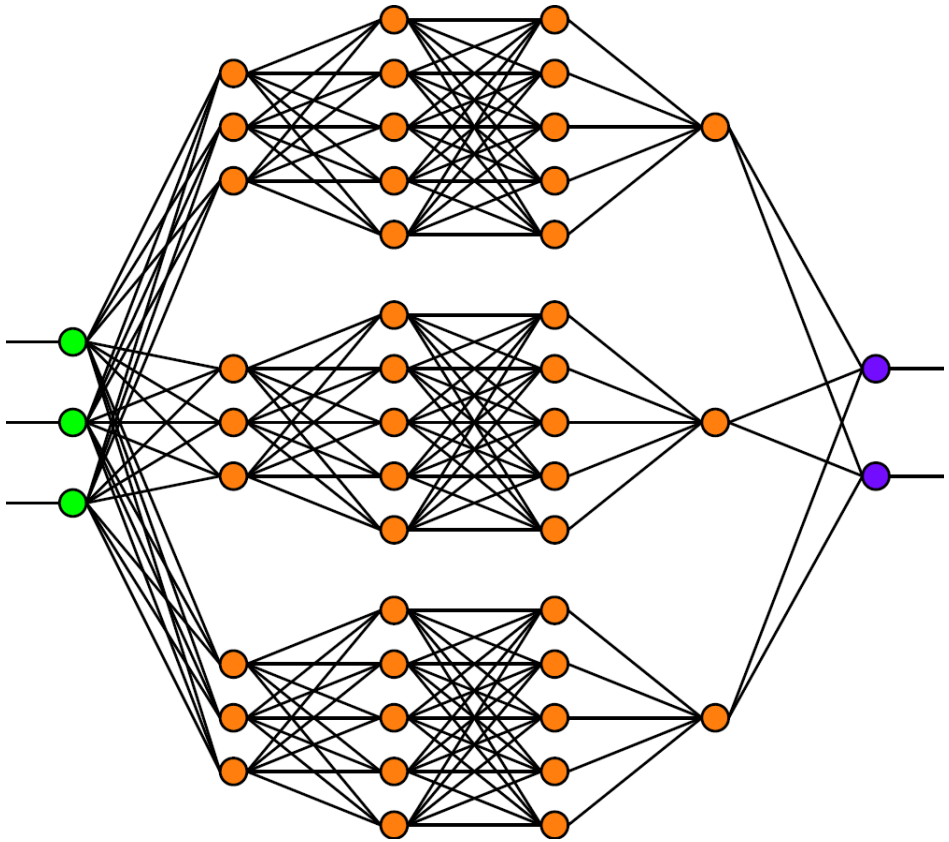
$$\|W_1^1\| \|W_1^2\| \|W_1^3\| \|W_1^4\| \|W_1^5\|$$



W_1^1 W_1^2 W_1^3 W_1^4 W_1^5

Regularizer Adapted to Network Size

- Sum over all subnetworks



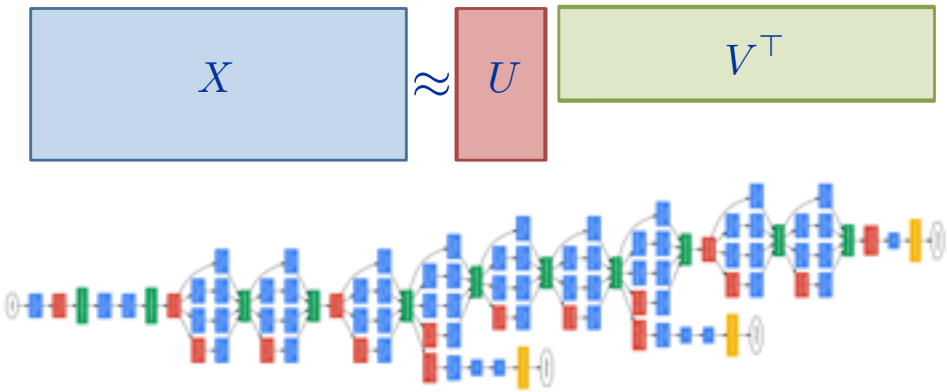
$$\Theta(W) = \sum_{i=1}^{r} \theta(W^i)$$

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

- 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.

Main Results: Matrix Factorization

- **Convex formulations**: $\min_X \ell(Y, X) + \lambda \|X\|_*$

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

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

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

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

- 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] Burer, Monteiro. Local minima and convergence in low- rank semidefinite programming. Math. Prog., 2005.
[2] Cabral, De la Torre, Costeira, Bernardino, "Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition," CVPR, 2013, pp. 2488–2495.
[3] Bach, Mairal, Ponce, Convex sparse matrix factorizations, arXiv 2008.
[4] Bach. Convex relaxations of structured matrix factorizations, arXiv 2013.

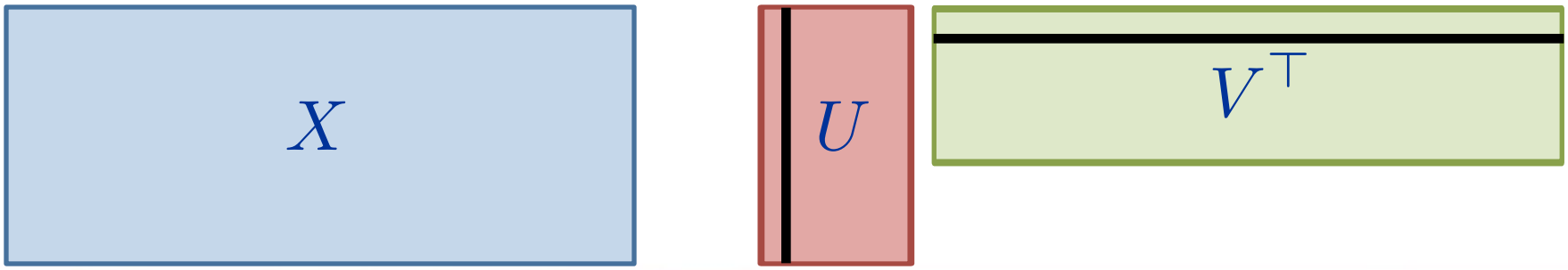
Main Results: Matrix Factorization

- **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

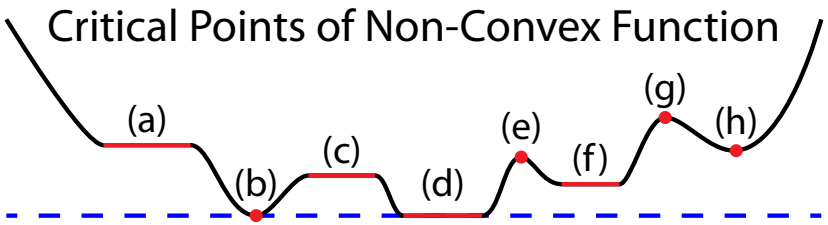
$$\min_{X} \ell(Y, X) + \lambda \|X\|_{u,v}$$

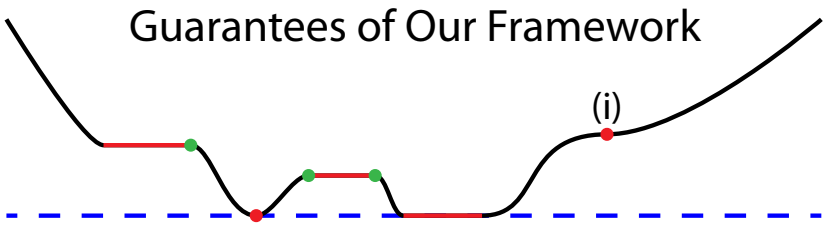


[1] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14
[2] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv '15

Main Results: Matrix Factorization

- **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: Tensor Fact. & Deep Learning

- In matrix factorization we had "generalized nuclear norm"

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

- By analogy we define "nuclear deep net regularizer"

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

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: Tensor Fact. & Deep Learning

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

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

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

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

[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.

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!