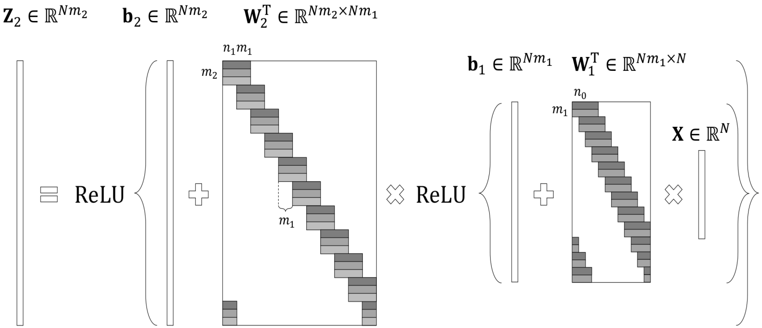
Outline for today

- A data model generated through a deep sparse deconvolutional model
- A notion of stripe sparsity based on locality of the features; stripe sparsity
- Proof that for such data the generating activations are obtained in a deep network formulation
- Examples of representations learned through the sparse deconvolutional models, and early results using ℓ^1 regularization.

CNN model through sparse coding (Papyan et al. 16'[1])

Consider a deep conv. net composed of two convolutional layers:

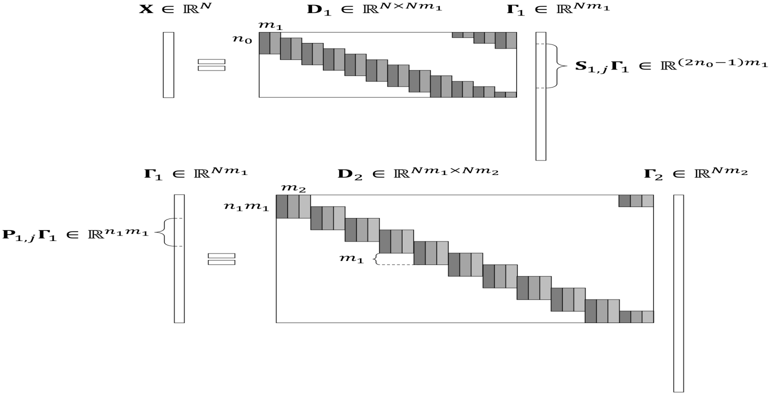


The forward map (note notation using transpose of $W^{(i)}$):

$$Z_2 = \sigma\left(b^{(2)} + (W^{(2)})^T \sigma\left(b^{(1)} + (W^{(1)})^T x\right)\right)$$

[1] https://arxiv.org/pdf/1607.08194.pdf

Deconvolutional NN data model (Papyan et al. 16'[2])



Two layer deconvolutional data model with weight matrices fixed, $W^{(i)} = D_i$, and $\Gamma_i \geq 0$ whose values compose data element X.

[2] https://arxiv.org/pdf/1607.08194.pdf

Stripe sparsity model (Papyan et al. 16'[3])

Consider a data vector x restricted to a patch of n consecutive entries, $x_i \in \mathbb{R}^n$. Due to the convolutional structure in D with m masks, each of length n, the portion of Γ that can influence x_i is the patch $\gamma_i \in \mathbb{R}^{(2n-1)m}$.

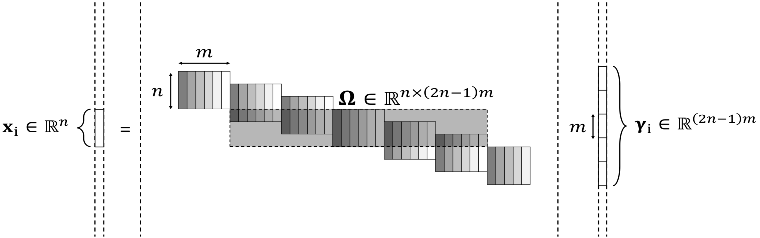


Figure 4: The i-th patch $\mathbf{x}_i$ of the global system $\mathbf{X} = \mathbf{D}\mathbf{\Gamma}$, given by $\mathbf{x}_i = \mathbf{\Omega}\boldsymbol{\gamma}_i$.

We consider Γ to have a stripe sparsity defined by $\|\Gamma\|^s_{0,\infty} = \max_i \|\gamma_i\|_0$.

[3] https://arxiv.org/pdf/1607.08194.pdf

Data model of union of subspaces (Papyan et al. 16'[4])

Consider the data model where for fixed known $\{D_i\}_{i=1}^{N}$ and stripe sparsity $\|\Gamma_i\|_{0,\infty}^{s} \leq s_i$ for $i = 1, \cdot, N$ the data is composed by

$$\begin{aligned} X &= D_1\Gamma_1 \\ \Gamma_1 &= D_2\Gamma_2 \\ &\vdots \\ \Gamma_{N-1} &= D_N\Gamma_N \end{aligned} \tag{1}$$

For such a data model is it guaranteed that a deep network with weights $W^{(i)} = D_i^T$ would have the same activations as Γ_i; that is would Γ_i be similar to $h_{i+1} = \sigma(W^{(i)}h_i)$ in some norm or otherwise?

[4] https://arxiv.org/pdf/1607.08194.pdf

Stability of layered hard thresholding (Papyan et al. 16'[5])

Theorem (Layered hard thresholding)

Let $Y = X + E$ where E denotes missfit to the model or noise and X be given by the data model (1null).

Let $\|E\|_{2,\infty}^{P} \leq \epsilon_0$ be a local bound on the error and let $\hat{\Gamma}_i = H_{\beta_i}\left(D_i^T \hat{\Gamma}_{i-1}\right)$ where $\hat{\Gamma}_0 = Y$, then if β_i are chosen appropriately (formulae available) and

$$\|\Gamma_i\|_{0,\infty}^{s} \leq \frac{1}{2}\left(1 + \mu^{-1}(D_i)\frac{|\Gamma_i^{min}|}{|\Gamma_i^{max}|}\right) - \mu^{-1}(D_i)\frac{\epsilon_{i-1}}{|\Gamma_i^{max}|}$$

then the support of $\hat{\Gamma}_i$ and Γ_i are the same and moreover $\|\Gamma_i - \hat{\Gamma}_i\|_{2,\infty}^{P} \leq \epsilon_i = \sqrt{\|\Gamma_i\|_{0,\infty}^{P}}\left(\epsilon_{i-1} + \mu(D_i)|\Gamma_i^{max}|(\|\Gamma_i\|_{0,\infty}^{s} - 1)\right)$.

For simple union of subspace data models the convolutional network is guaranteed to recover the generating activations with

Learned ML-CSC on MNIST (Sulam et al. 18'[6])

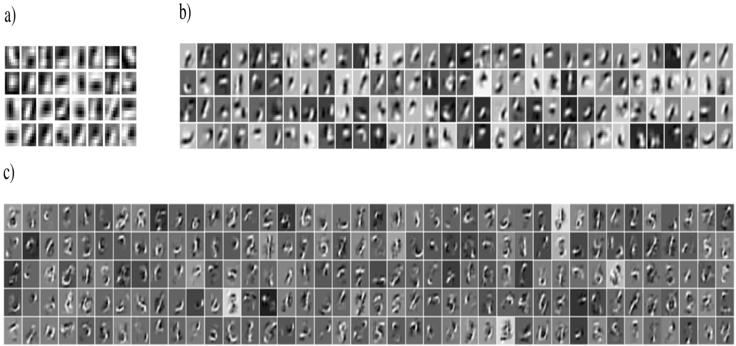


Fig. 3: ML-CSC model trained on the MNIST dataset. a) The local filters of the dictionary D_1. b) The local filters of the effective dictionary $D^{(2)} = D_1 D_2$. c) Some of the 1024 local atoms of the effective dictionary $D^{(3)}$ which, because of the dimensions of the filters and the strides, are global atoms of size 28×28.

Learned dictionaries are show increasing structure from local wavelets in D_1 to composite features in D_2 to representative numbers in D_3.

[6] https://arxiv.org/abs/1708.08705

Stability of layered ℓ^1-regularization (Papyan et al. 16'[7])

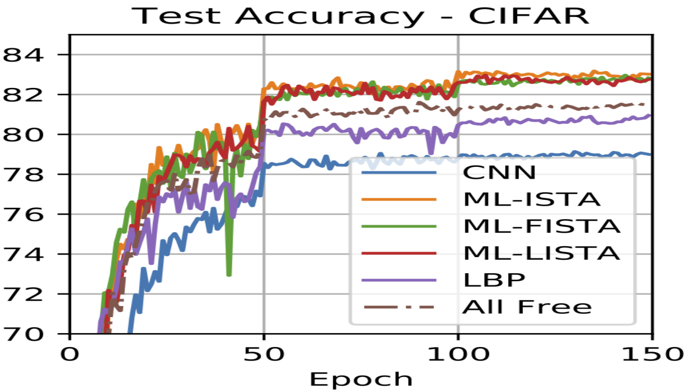
Theorem (Layered ℓ^1-regularization)

Let $Y = X + E$ where E denotes missfit to the model or noise and X be given by the data model (1null).
Let $\|E\|_{2,\infty}^P \leq \epsilon_0$ be a local bound on the error and let $\hat{\Gamma}_i = \text{argmin}_\Gamma \xi_i \|\Gamma\|_1 + \frac{1}{2}\|D_i\Gamma - \hat{\Gamma}_{i-1}\|_2^2$ where $\hat{\Gamma}_0 = Y$, then if $\xi_i = 4\epsilon_{i-1}$ and $\|\Gamma_i\|_{0,\infty}^s \leq \frac{1}{3}\left(1 + \mu^{-1}(D_i)\right)$ then the support of $\hat{\Gamma}_i$ and Γ_i are the same and moreover
$\|\Gamma_i - \hat{\Gamma}_i\|_{2,\infty}^P \leq \epsilon_i = \|E\|_{2,\infty}^P 7.5^i \Pi_{j=1}^i \sqrt{\|\Gamma_j\|_{0,\infty}^P}$.

More complex methods to determine activations give provable recovery with less strict conditions, here $\|\Gamma_i\|_{0,\infty}^s$ has not dependence on the magnitude of entries.

[7] https://arxiv.org/pdf/1607.08194.pdf

Accuracy of multi-layer ℓ^1-regularizers (Sulam et al. 18'[8])



Three layer networks with ℓ^1 regularization through (F)ISTA vs. a six layer CNN (three layers convolutional layers followed by fully connected layers). LISTA and LBP are variants also using ℓ^1 regularization.

[8] https://arxiv.org/pdf/1806.00701.pdf

Union of randomized subspaces (Murray et al. 18'[9])

Consider the data model where for fixed known $\{D_i\}_{i=1}^{N}$ and stripe sparsity $\|\Gamma_i\|_{0,\infty}^{s} \leq s_i$ for $i = 1, \cdot, N$ the data is composed by

$$\begin{aligned} X &= D_1\Sigma_1\Gamma_1 + V_0 \\ \Gamma_1 &= D_2\Sigma_2\Gamma_2 + V_1 \\ &\vdots \\ \Gamma_{N-1} &= D_N\Sigma_N\Gamma_N + V_{N-1} \end{aligned} \quad (2)$$

where Σ_i are diagonal matrices whose diagonal is composed of randomly drawn ± 1.

Introducing this randomness allows us to further weaken the conditions on the coherence needed to guarantee recovery.

[9] https://ieeexplore.ieee.org/document/8439894

Provable activation pathway recovery (Murray et al. 18'[10])

Theorem hard thresholding)

Let $\hat{\mathbf{X}}^{(l-1)}$ be consistent with the D-CSC model (2null), with $\|\mathbf{V}^{(l)}\|_{2,\infty}^{P^{(l)}} \leq \zeta_l$ and $\|\mathbf{X}^{(l)}\|_{0,\infty}^{Q^{(l)}} \leq S_l$ for all $l = 0, \ldots, L-1$, and $\Sigma^{(l)}$ diagonal matrices with independent Rademacher random variables. Let denote as Z_L the event that the activation path is successfully recovered by hard thresholding $\hat{\Gamma}_i = H_{S_i}\left(D_i^T \hat{\Gamma}_{i-1}\right)$. Then

$$P(\bar{Z}_L) \leq 2dM \sum_{l=1}^{L} n_l \exp\left(-\frac{|X_{min}^{(l)}|^2}{8\left(|X_{max}^{(l)}|^2 \mu_l^2 S_l + \zeta_{l-1}^2\right)}\right).$$

Where $X^{(0)} \in \mathbb{R}^{M \times d}$ and filters at layer l are of length n_l.

The derived probability bound scales proportional to μ_l^{-2} across a given layer, rather than μ_l^{-1}

[10] https://ieeexplore.ieee.org/document/8439894

One step thresholding: average sign pattern [ScVa07]

Input: y, D and k (number of nonzeros in output vector).
Algorithm: Set Λ the index set of the $k \leq m$ largest in $|D^* y|$
Output the n-vector x whose entries are

$$x_\Lambda = (D_\Lambda^* D_\Lambda)^{-1} D_\Lambda y \quad \text{and} \quad x(i) = 0 \text{ for } i \notin \Lambda.$$

One step thresholding: average sign pattern [ScVa07]

Input: y, D and k (number of nonzeros in output vector).
Algorithm: Set Λ the index set of the $k \leq m$ largest in $|D^* y|$
Output the n-vector x whose entries are

$$x_\Lambda = (D_\Lambda^* D_\Lambda)^{-1} D_\Lambda y \quad \text{and} \quad x(i) = 0 \text{ for } i \notin \Lambda.$$

Theorem

Let $y = Dx_0$, with the columns of D having unit ℓ^2 norm, the sign of the nonzeros in x_0 selected randomly from ± 1 independent of D, and

$$\|x_0\|_{\ell^0} < (128 \log(2n/\epsilon))^{-1} \nu_\infty^2(x_0) \mu_2^{-2}(D),$$

then, with probability greater than $1 - \epsilon$, the Thresholding decoder with $k = \|x_0\|_{\ell^0}$ will return x_0.

One step thresholding: average sign pattern (proof, pg. 1)

Theorem (Rademacher concentration)

Fix a vector α. Let ϵ be a Rademacher series, vector with entries drawn uniformly from ± 1, of the same length as α, then

$$\text{Prob}\left(\left|\sum_i \epsilon_i \alpha_i\right| > t\right) \leq 2\exp\left(\frac{-t^2}{32\|\alpha\|_2^2}\right)$$

Let $\Lambda := \text{supp}(x_0)$. Thresholding fail to recover x_0 if

$$\max_{i \notin \Lambda} |d_i^* y| > \min_{i \in \Lambda} |d_i^* y|.$$

$$\text{Prob}\left(\max_{i \notin \Lambda} |d_i^* y| > p \quad \text{and} \quad \min_{i \in \Lambda} |d_i^* y| < p\right) \quad \leq$$

$$\text{Prob}\left(\max_{i \notin \Lambda} |d_i^* y| > p\right) + \text{Prob}\left(\min_{i \in \Lambda} |d_i^* y| < p\right) \quad =: \quad P_1 + P_2$$

One step thresholding: average sign pattern (proof, pg. 2)

$$\begin{aligned}
P_1 &= \text{Prob}\left(\max_{i\notin\Lambda}|d_i^* y| > p\right) \\
&\leq \sum_{i\notin\Lambda} \text{Prob}\left(|d_i^* y| > p\right) \\
&= \sum_{i\notin\Lambda} \text{Prob}\left(\left|\sum_{j\in\Lambda} x_0(j)(d_i^* d_j)\right| > p\right) \\
&\leq 2\sum_{i\notin\Lambda} \exp\left(\frac{-p^2}{32\sum_{j\in\Lambda}|x_0(j)|^2|d_i^* d_j|^2}\right) \\
&\leq 2(n-k)\exp\left(\frac{-p^2}{32k\|x_0\|_\infty^2\mu_2^2(D)}\right).
\end{aligned}$$

One step thresholding: average sign pattern (proof, pg. 3)

$$
\begin{aligned}
P_2 &= \text{Prob}\left(\min_{i\in\Lambda} |d_i^* y| < p\right) \\
&\leq \text{Prob}\left(\min_{i\in\Lambda} |x_0(i)| - \max_{i\in\Lambda}\left|\sum_{j\in\Lambda, j\neq i} x_0(j)(d_i^* d_j)\right| < p\right) \\
&\leq \sum_{i\in\Lambda} \text{Prob}\left(\left|\sum_{j\in\Lambda, j\neq i} x_0(j)(d_i^* d_j)\right| > \min_{i\in\Lambda} |x_0(i)| - p\right) \\
&\leq 2\sum_{i\in\Lambda} \exp\left(\frac{-(\min_{i\in\Lambda} |x_0(i)| - p)^2}{32\sum_{j\in\Lambda, j\neq i} |x_0(j)|^2 |d_i^* d_j|^2}\right) \\
&\leq 2k \exp\left(\frac{-(\min_{i\in\Lambda} |x_0(i)| - p)^2}{32k\|x_0\|_\infty^2 \mu_2^2(D)}\right).
\end{aligned}
$$

One step thresholding: average sign pattern (proof, pg. 4)

Balance P_1 and P_2 by setting $p := \min_{i\in\Lambda} |x_0(i)|/2$:

$$P_1 + P_2 \leq 2n \exp\left(\frac{-(\min_{i\in\Lambda} |x_0(i)|)^2}{128k\|x_0\|_\infty^2 \mu_2^2(D)}\right) \leq 2n \exp\left(\frac{-\nu_\infty(x_0)^2}{128k\mu_2^2(D)}\right).$$

Setting this bound on the probability of failure equal to ϵ and solving for k yields the conclusion of the proof. □

- Similar work for matching pursuit by Schnass, ℓ^1 by Tropp, and in Statistical RICs
- Stronger uniform statements we need more than coherence.

Deep convolutional sparse coding: summary

- By constructing a union of subspace data model we can employ methods of analysis developed by the compressed sensing community.
- Data of this type provably have the activations one would expect based on the data construction.
- Recovery is possible for nonlinear activations which include: soft or hard thresholding as well as ℓ^1-regularization.
- The data model isn't as rich as we would hope as it is linear
- Recovery guarantees scale poorly with depth and are based on coherence between filters which are not small for local convolutional filters; recall Grassmann frame bounds.
- Open questions include the role of activations, learning the features, and building in structure within and between labels.