On Stability of Deep Structured Linear Network

Pradip Sasmal

IISc Bangalore

02/11/2019

Multilinear Compressed Sensing

Goal

Aim is to evaluate how far the deep linear network architectures used in applications are from architectures for which one can guarantee that the parameters returned by the algorithm, and therefore the features defined using these parameters, are stably defined.

Results

- Derived a necessary and sufficient conditions on the network topology under which a stability property holds.
- The stability property requires that the error on the parameters defining the near-optimal factors scales linearly with the reconstruction error (i.e., the risk).
- Under stability conditions on the network topology, any successful learning task leads to stably defined features that can be interpreted.

The Road Map

- First evaluate how the Segre embedding and its inverse distort distances.
- Any deep structured linear network can be cast as a generic multilinear problem that uses the Segre embedding. This is the tensorial lifting.
- Using the tensorial lifting, a necessary and sufficient condition for the identifiability of the factors up to a scale rearrangement.
- Finally provide a necessary and sufficient condition called the deep-Null Space Property (because of the analogy with the usual Null Space Property in the compressed sensing framework) which guarantees that the stability property holds.

Informal Statement of the Main Theorem

Assume a known parameterized family of functions $f_{\mathbf{h}}$ and a metric d between parameter pairs. A necessary and sufficient condition on the family $f_{\mathbf{h}}$ guaranteeing the following:
There exists a constant $C > 0$ such that for any input/output pairs I, X and any pair of parameters $\mathbf{h}^*, \overline{\mathbf{h}}$ for which

$$\delta = \|X - f_{\mathbf{h}^*}(I)\|$$

and

$$\eta = \left\|X - f_{\overline{\mathbf{h}}}(I)\right\|$$

are sufficiently small, one have

$$d\left(\bar{h}, h^*\right) \leq C(\delta + \eta) \tag{1}$$

- Inequality (1) therefore guarantees that the set made of the parameters leading to a small risk has a small diameter. The features defined using such parameters are therefore stably defined.

Deep Structured Linear Networks

- Consider an arbitrary depth parameter $K \geq 1$. The number of layers is $K+1$, and the layer receiving the input is $K+1$ and the layer returning the output is 1.
- Let $m_1 \dots m_{K+1} \in \mathbb{N}$ s.t. $m_1 = m$ and $m_{K+1} = n$. For $k = 1, \dots, K$, consider the linear map

$$\begin{aligned} M_k : \mathbb{R}^S &\longrightarrow \mathbb{R}^{m_k \times m_{k+1}} \\ h &\longmapsto M_k(h) \end{aligned}$$

- Given some parameters, $h_1, \dots, h_K \in \mathbb{R}^S$, the action of the deep structured linear network is the product $M_1(h_1) \cdots M_K(h_K)$.
- $M_K(h_K) = M'_K(h_K)\, I$ for a linear map M'_K and for a matrix I whose columns contain the inputs.
- Given outputs $X \in \mathbb{R}^{m \times n}$, the optimization of the parameters $h_1, \dots, h_K$ defining the network aims at getting

$$M_1(h_1) \cdots M_K(h_K) \simeq X$$

Structure on M_K

- To model feed-forward linear networks, the mappings $M_k, k = 1, \ldots, K-1$ (and M'_K) construct the matrix by placing the entry of h_k corresponding to an edge in the network in the corresponding entry in $M_k(h_k)$.
- For convolutional layers, M_k and M'_K concatenate convolution matrices' defined by a portion of the entries in $\mathbf{h}_k$. Each convolution matrix is at the location corresponding to a prescribed edge.
- The main argument for studying deep structured linear networks is due to their strong connection to nonlinear networks that uses the rectified linear unit (ReLU) " activation function. We explain it in detail. The action of the ReLU activation function at the layer k treats every entry independently of the other entries and multiplies it by either 1 (the entry is kept) or 0 (the entry is canceled).

Structure on $\mathbf{h}$

- In addition to the structure induced by the operators M_k, we also consider a structure imposed on the vectors h We assume that we know a collection of models $\mathcal{M} = \left(\mathcal{M}^L\right)_{L \in \mathbb{N}}$ with the property that for every $L, \mathcal{M}^L \subset \mathbb{R}^{S \times K}$ is a given subset. We will assume that the parameters $\mathbf{h} \in \mathbb{R}^{S \times K}$ defining the factors are such that there exists $L \in \mathbb{N}$ such that $\mathbf{h} \in \mathcal{M}^L$.
- For instance, the constraint $\mathbf{h} \in \mathcal{M}^L$ might be used to impose sparsity, grouped sparsity, or cosparsity. One might also use the constraint $\mathbf{h} \in \mathcal{M}^L$ to impose nonnegativity, orthogonality, equality, compactness, etc. Generally speaking $\mathcal{M}$ is used to impose some prior or some form of regularity or to compress the parameter space and obtain better bounds.

Matrix Factorization and Compressed Sensing

- In signal processing, one usually know that $\bar{h}$ exists and δ represents the sum of a modeling error and noise. Inequality (1) guarantees that when the condition is satisfied, even an approximative minimizer of

$$\text{argmin}_{L\in\mathbb{N},(h_k)_{k=1..K}\in\mathcal{M}^L} \|M_1(h_1)\cdots M_K(h_K) - X\|^2 \quad (2)$$

  leads to a solution h^* close $\bar{h}$. This property is often named the stable recovery guarantee.
- When $\delta = 0$ (i.e., the data exactly fits the model and is not noisy) and $\eta = 0$ (i.e., (2) is perfectly solved) this identifiability guarantee. This is a necessary condition of stable recovery.
- $K = 1$: Linear inverse problems, $K = 2$: Bilinear inverse problems and bilinear parameterizations (For example: Nonnegative matrix factorization and low rank prior, Phase retrieval, Sparse coding and dictionary learning)

Notations

- $T \in \mathbb{R}^{S^K}$: real-valued tensors of order K whose axes are of size S for $K \geq 1$ and $S \geq 2$, $\mathbb{R}^{S^K}$: space of tensors, $T_{i_1 \ldots i_K}$: The entries of T, where $(i_1, \ldots, i_K) \in [S]^K$.
- $\mathbf{h} \in \mathbb{R}^{S \times K}$: the parameters defining the factors are gathered in a single matrix, $\mathbf{h}_k \in \mathbb{R}^S$: The k th vector containing the parameters for the layer k, $\mathbf{h}_{k,i} \in \mathbb{R}$: The i th entry of the k th vector $h \in \mathbb{R}^S$: A vector not related to an element in $\mathbb{R}^{S \times K}$
- $\mathcal{M} = \left(\mathcal{M}^L\right)_{L \in \mathbb{N}}$, with $\mathcal{M}^L \subset \mathbb{R}^{S \times K}$
- $\mathbf{h} \in \mathbb{R}^{S \times K}$ and $T \in \mathbb{R}^{S^K}$, $p < +\infty$
  $\|\mathbf{h}\|_p = \left(\sum_{k=1}^{K} \sum_{i=1}^{S} |\mathbf{h}_{k,i}|^p\right)^{1/p}$, $\|T\|_p = \left(\sum_{\mathbf{i} \in [S]^K} |T_{\mathbf{i}}|^p\right)^{1/p}$
- $\|\mathbf{h}\|_\infty = \max_{\substack{k \in [K] \\ i \in [S]}} |\mathbf{h}_{k,i}|$, $\quad \|T\|_\infty = \max_{\mathbf{i} \in [S]^K} |T_{\mathbf{i}}|$

Continue ...

- Set $\mathbb{R}_*^{S\times K} = \left\{\mathbf{h} \in \mathbb{R}^{S\times K} | \forall k \in [K], \|\mathbf{h}_k\| \neq 0\right\}$
- Define an equivalence relation on $\mathbb{R}_*^{S\times K}$: For any $\mathbf{h}, \mathbf{g} \in \mathbb{R}^{S\times K}, \mathbf{h} \sim \mathbf{g}$ if and only if there exist $(\lambda_k)_{k\in[K]} \in \mathbb{R}^K$ such that

$$\prod_{k=1}^{K} \lambda_k = 1 \quad \text{and} \quad \mathbf{h}_k = \lambda_k \mathbf{g}_k \quad \forall k \in [K]$$

Denote the equivalence class of $\mathbf{h} \in \mathbb{R}_*^{S\times K}$ by $\langle \mathbf{h} \rangle$

Rank of a Tensor

- The zero tensor is of rank 0. A nonzero tensor $T \in \mathbb{R}^{S^K}$ is of rank 1 (or decomposable) if and only if there exists $\mathbf{h} \in \mathbb{R}_*^{S \times K}$ such that T is the outer product of the vectors $\mathbf{h}_k$ for $k \in [K]$. That is, for any $\mathbf{i} \in [S]^K$,

$$T_{\mathbf{i}} = \mathbf{h}_{1,i_1} \cdots \mathbf{h}_{K,i_K}.$$

  Let $\Sigma_1 \subset \mathbb{R}^{S^K}$ denote the set of tensors of rank 0 or 1.
- The rank of a tensor $T \in \mathbb{R}^{S^K}$ is

$$\text{rk}\,(T) = \min\{r \in \mathbb{N} | \exists T_1, \ldots, T_r \in \Sigma_1 \quad \text{s.t.} \;\; T = T_1 + \cdots + T_r\}$$

- For $r \in \mathbb{N}$, let $\Sigma_r = \left\{ T \in \mathbb{R}^{S^K} \mid \text{rk}(T) \leq r \right\}$

Segre embedding and tensors of rank 1 and 2

$$P : \mathbb{R}^{S\times K} \longrightarrow \Sigma_1 \subset \mathbb{R}^{S^K}$$

$$\mathbf{h} \longmapsto (\mathbf{h}_{1,i_1}\mathbf{h}_{2,i_2}\cdots\mathbf{h}_{K,i_K})_{\mathbf{i}\in[S]^K}$$

The map P is called the Segre embedding and is often denoted $\widehat{Seg}$ in the algebraic geometry literature.

- Identifiability of $\langle\mathbf{h}\rangle$ from $P(\mathbf{h})$: For $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}_*^{S\times K}$, $P(\mathbf{h}) = P(\mathbf{g})$ if and only if $\langle\mathbf{h}\rangle = \langle\mathbf{g}\rangle$
- Geometrical description of $\Sigma_{1,*}$: $\Sigma_{1,*}$ is a smooth (i.e., C^∞) manifold of dimension $K(S-1)+1$
- $\mathbb{R}^{S\times K}_{\text{diag}} = \left\{\mathbf{h} \in \mathbb{R}_*^{S\times K} | \forall k \in [K], \|\mathbf{h}_k\|_\infty = \|\mathbf{h}_1\|_\infty\right\}$
  For any $p \in [1,\infty]$, define the mapping
  $d_p : \left(\mathbb{R}_*^{S\times K}/\sim \times \mathbb{R}_*^{S\times K}/\sim\right) \to \mathbb{R}$ by

$$d_p(\langle\mathbf{h}\rangle,\langle\mathbf{g}\rangle) = \inf_{\mathbf{h}'\in\langle\mathbf{h}\rangle\cap\mathbb{R}^{S\times K}_{diag},\mathbf{g}'\in\langle\mathbf{g}\rangle\cap\mathbb{R}^{S\times K}_{diag}} \left\|\mathbf{h}'-\mathbf{g}'\right\|_p \quad \forall \mathbf{h},\mathbf{g} \in \mathbb{R}_*^{S\times K}$$

- For any $p \in [1,\infty]$, d_p is a metric on $\mathbb{R}_*^{S\times K}/\sim$

Continue ...

Theorem

(stability of $\langle \mathrm{h} \rangle$ from $P(\mathrm{h})$) Let h and $\mathrm{g} \in \mathbb{R}_*^{S \times K}$ be such that $\|P(\mathrm{g}) - P(\mathrm{h})\|_\infty \leq \frac{1}{2} \max\left(\|P(\mathrm{h})\|_\infty, \|P(\mathrm{g})\|_\infty\right)$. Then for all $p, q \in [1, \infty]$,

$$d_p(\langle \mathbf{h} \rangle, \langle \mathbf{g} \rangle) \leq 7(KS)^{\frac{1}{p}} \min\left(\|P(\mathbf{h})\|_\infty^{\frac{1}{K}-1}, \|P(\mathbf{g})\|_\infty^{\frac{1}{K}-1}\right) \|P(\mathbf{h}) - P(\mathbf{g})\|_q$$

Theorem

("Lipschitz continuity"" of P) For any $q \in [1, \infty]$ and any h and $\mathbf{g} \in \mathbb{R}_*^{S \times K}$,

$$\|P(\mathbf{h}) - P(\mathbf{g})\|_q \leq S^{\frac{K-1}{q}} K^{1-\frac{1}{q}} \max\left(\|P(\mathbf{h})\|_\infty^{1-\frac{1}{K}}, \|P(\mathbf{g})\|_\infty^{1-\frac{1}{K}}\right) d_q(\langle \mathbf{h} \rangle, \langle \mathbf{g} \rangle)$$

The Tensorial Lifting

- Let $M_k, k \in [K]$, be as in (1). The entries of the matrix

$$M_1(\mathbf{h}_1) M_2(\mathbf{h}_2) \cdots M_K(\mathbf{h}_K)$$

are multivariate polynomials whose variables are the entries of $\mathbf{h} \in \mathbb{R}^{S \times K}$. Moreover, every entry is the sum of monomials of degree K. Each monomial is a constant times $\mathbf{h}_{1,i,i} \cdots \mathbf{h}_{K,i_K}$ for some $\mathbf{i} \in [S]^K$.

- (tensorial lifting). Let $M_k, k \in [K]$ be as in (1.2). The map

$$(\mathbf{h}_1, \ldots, \mathbf{h}_K) \longmapsto M_1(\mathbf{h}_1) M_2(\mathbf{h}_2) \cdots M_K(\mathbf{h}_K)$$

uniquely determines a linear map

$$\mathcal{A} : \mathbb{R}^{S^K} \longrightarrow \mathbb{R}^{m \times n}$$

such that for all $h \in \mathbb{R}^{\times K}$,

$$M_1(\mathbf{h}_1) M_2(\mathbf{h}_2) \cdots M_K(\mathbf{h}_K) = \mathcal{A}P(\mathbf{h}). \tag{3}$$

We call (3) and its use the tensorial lifting.

Continue ...

Using (3) when (2) has a minimizer, we rewrite it in the form

$$\mathbf{h}^* \in \operatorname{argmin}_{L\in\mathbb{N},\mathbf{h}\in\mathcal{M}^L} \|\mathcal{A}P(\mathbf{h}) - X\|^2 \tag{4}$$

We now decompose this problem into two subproblems: a least-squares problem,

$$T^* \in \operatorname{argmin}_{T\in\mathbb{R}^{SK}} \|\mathcal{A}T - X\|^2 \tag{5}$$

and a nonconvex problem,

$$\mathbf{h}'^* \in \operatorname{argmin}_{L\in\mathbb{N},\mathbf{h}\in\mathcal{M}^L} \|\mathcal{A}\left(P(\mathbf{h}) - T^*\right)\|^2 \tag{6}$$

Let X and $\mathcal{A}$ be such that (2) has a minimizer:

- Let $\mathbf{h}^*$ be a solution of (4) Then, for any solution T^* of (5) $\mathbf{h}^*$ also minimizes (6)
- Let T^* be a solution of (5) and $\mathbf{h}'^*$ solution of (6). Then $\mathbf{h}'^*$ also minimizes (4).

Identifiability (error-free case)

Assume that X is such that there exist $\bar{L}$ and $\overline{\mathbf{h}} \in \mathcal{M}^{\bar{L}}$ such that

$$X = M_1\left(\overline{\mathbf{h}}_1\right) \cdots M_K\left(\overline{\mathbf{h}}_K\right) \tag{7}$$

Under this assumption, $X = \mathcal{A}P(\overline{\mathbf{h}})$, so

$$P(\overline{\mathbf{h}}) \in \operatorname{argmin}_{T \in \mathbb{R}^{sK}} \|\mathcal{A}T - X\|^2$$

Moreover, we trivially have $P(\overline{\mathbf{h}}) \in \Sigma_1$, and therefore $\overline{\mathbf{h}}$ minimizes (6) and As a consequence,(2) has a minimizer.

Definition

(identifiability). We say that $\langle\overline{\mathbf{h}}\rangle$ is identifiable if the elements of $\langle\overline{\mathbf{h}}\rangle$ are the only solutions of (2.1).

We say that $\mathcal{M}$ identifiable if for every $L \in \mathbb{N}$ and every $\overline{\mathbf{h}} \in \mathcal{M}^L$, $\langle\overline{\mathbf{h}}\rangle$ is identifiable.

Continue ...

- (characterization of the global minimizers) For any $L^* \in \mathbb{N}$ and any h'$\in \mathrm{M}^{L^*}$, $(L^*, \mathbf{h}^*) \in \operatorname{argmin}_{L\in\mathbb{N},\mathbf{h}\in\mathcal{M}} \|\mathcal{A}P(\mathbf{h}) - X\|^2$ if and only if
$$P(\mathbf{h}^*) \in P(\bar{\mathbf{h}}) + \mathrm{Ker}(\mathcal{A})$$
- Define for any $L' \in \mathbb{N}$
$$P\left(\mathcal{M}^L\right) - P\left(\mathcal{M}^{L'}\right) := \left\{P(\mathbf{h}) - P(\mathbf{g}) |\, \mathbf{h} \in \mathcal{M}^L, \mathbf{g} \in \mathcal{M}^{L'}\right\} \subset \mathbb{R}^{S^K}$$
- (necessary and sufficient conditions of identifiability).
1. For any $\bar{L}$ and $\bar{\mathbf{h}} \in \mathcal{M}^{\bar{L}}$, $\langle\bar{\mathbf{h}}\rangle$ is identifiable if and only if for any $L \in \mathbb{N}$ $(P(\bar{\mathbf{h}}) + \mathrm{Ker}(\mathcal{A})) \cap P\left(\mathcal{M}^L\right)$ C $\{P(\bar{\mathbf{h}})\}$.
2. $\mathcal{M}$ is identifiable if and only if for any L and $L' \in \mathbb{N}$
$\mathrm{Ker}\,(\mathcal{A}) \cap \left(P\left(\mathcal{M}^L\right) - P\left(\mathcal{M}^{L'}\right)\right) \subset \{0\}$

Stability guarantee

- Assume that there exist $\bar{L}$ and $L^* \in \mathbb{N}, \overline{\mathbf{h}} \in \mathcal{M}^{\bar{L}}$, and $\mathbf{h}^* \in \mathcal{M}^{L^*}$, such that

$$\| M_1\left(\overline{\mathbf{h}}_1\right) \cdots M_K\left(\overline{\mathbf{h}}_K\right) - X \leq \delta \tag{8}$$

and

$$\| M_1\left(\mathbf{h}_1^*\right) \cdots M_K\left(\mathbf{h}_K^*\right) - X \leq \eta \tag{9}$$

for δ and η typically small.

Definition

(deep-Null Space Property). Let $\gamma > 0$ and $\rho > 0$. We say that Ker $(\mathcal{A})$ satisfies the deep-Null Space Property (deep-NSP) with respect to the collection of models $\mathcal{M}$ with constants (γ, ρ) if for any L and $L' \in \mathbb{N}$, any $T \in P\left(\mathcal{M}^L\right) - P\left(\mathcal{M}^{L'}\right)$ satisfying $\|\mathcal{A}T\| \leq \rho$ and any $T' \in \text{Ker}(\mathcal{A})$, we have $\|T\| \leq \gamma \|T - T'\|$.

Implication of deep-NSP

The deep-NSP implies that, for $T \in P\left(\mathcal{M}^L\right) - P\left(\mathcal{M}^{L'}\right)$ close to $\mathrm{Ker}(\mathcal{A})$ in the sense that $\|\mathcal{A}T\| \leq \rho$, we must have, by decomposing $T = T' + T''$, with $T' \in \mathrm{Ker}(\mathcal{A})$ and T'' in its orthogonal complement,

$$\|T\| \leq \gamma \left\|T - T'\right\| = \gamma \left\|T''\right\| \leq \frac{\gamma}{\sigma_{\min}} \left\|\mathcal{A}T''\right\| \leq \frac{\gamma}{\sigma_{\min}} \rho$$

where σ_{min} is the smallest nonzero singular value of $\mathcal{A}$. In words, $\|T\|$ must be small. We can conclude that under the deep-NSP, $P\left(\mathcal{M}^L\right) - P\left(\mathcal{M}^{L'}\right)$ and $\left\{T \in \mathbb{R}^{s^K} | \|\mathcal{A}T\| \leq \rho\right\}$ intersect at most in the neighborhood of 0.

Example

- If $\mathrm{Ker}(\mathcal{A}) = \{0\}$, then $\mathrm{Ker}(\mathcal{A})$ satisfies the deep-NSP with respect to the model $\mathbb{R}^{S\times K}$ with constants $(1, +\infty)$.
- For any $\gamma' \geq \gamma$: If Ker $(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}$ with constants (γ, ρ), then $\mathrm{Ker}(\mathcal{A})$ satisfies the deep-NSP with respect tothe collection of models $\mathcal{M}$ with constants (γ', ρ) .
- For any $\mathcal{M}' \subset \mathcal{M}$: If $\mathrm{Ker}(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}$ with constants (γ, ρ), then $\mathrm{Ker}(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}'$ with constants (γ, ρ).
- In particular, if $\mathrm{Ker}(\mathcal{A})$ satisfies the deep-NSP with respect to the model $\mathbb{R}^{S\times K}$ with constants (γ, ρ), it satisfies the deep-NSP with respect to any collection of models, with constants (γ, ρ).

Sufficient condition for the stability property

Assume Ker $(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}$ and with the constants (γ, ρ). For any $\mathbf{h}^*$ as in (9) with η and δ such that $\delta + \eta \leq \rho$, we have

$$\left\| P\left(\mathbf{h}^*\right) - P(\overline{\mathbf{h}}) \right\| \leq \frac{\gamma}{\sigma_{\min}} (\delta + \eta)$$

where $\sigma_{\min}$ is the smallest nonzero singular value of $\mathcal{A}$. Moreover, if $\overline{\mathbf{h}} \in \mathbb{R}_*^{S \times K}$ and $\frac{\gamma}{\sigma_{\min}}(\delta + \eta) \leq \frac{1}{2} \max\left(\|P(\overline{\mathbf{h}})\|_\infty, \|P\left(\mathbf{h}^*\right)\|_\infty \right)$, then

$$d_p\left(\langle \mathbf{h}^* \rangle, \langle \overline{\mathbf{h}} \rangle \right) \leq \frac{7(KS)^{\frac{1}{p}} \gamma}{\sigma_{\min}} \min\left(\|P(\overline{\mathbf{h}})\|_\infty^{\frac{1}{K}}, \|P\left(\mathbf{h}^*\right)\|_\infty^{\frac{1}{K}-1} \right) (\delta + \eta) \tag{10}$$

Necessary condition for the stability property

Assume the stability property holds: There exist C and $\delta > 0$ such that for any $\bar{L} \in \mathbb{N}, \overline{\mathbf{h}} \in \mathcal{M}^{\bar{L}}$, any $X = \mathcal{A}\mathcal{P}(\overline{\mathbf{h}}) + e$, with $\|e\| \leq \delta$, any $L^* \in \mathbb{N}$, and any h' $^* \in \mathcal{M}^{L^*}$ such that

$$\left\|\mathcal{A}\mathcal{P}\left(\mathbf{h}^*\right) - X\right\|^2 \leq \|e\|$$

we have

$$d_2\left(\left\langle \mathbf{h}^* \right\rangle, \left\langle \overline{\mathbf{h}} \right\rangle\right) \leq C \min\left(\|P(\overline{\mathbf{h}})\|_\infty^{\frac{1}{K}-1}, \|P\left(\mathbf{h}^*\right)\|_\infty^{\frac{1}{K}-1}\right)\|e\|$$

Then Ker $(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}$ with constants

$$(\gamma, \rho) = \left(CS^{\frac{K-1}{2}}\sqrt{K}\sigma_{\max}, \delta\right)$$

where $\sigma_{\max}$ is the spectral radius of $\mathcal{A}$.

Reference

F. Malgouyres, and J. Landsberg. “Multilinear compressive sensing and an application to convolutional linear networks.” SIAM Journal on Mathematics of Data Science 1.3 (2019): 446-475.

THANK YOU!