

# Another polynomial-time estimation algorithm

Miaoyan Wang, July 8, 2021

## 1 Rank- $\sqrt{d}$ approximatable

**Definition 1** (Rank- $\sqrt{d}$  approximatable tensor). Let  $\Theta$  be an order-3 tensor. We use  $f: [d] \rightarrow \mathbb{R}$  to denote the distance function, in the sense of matrix spectral norm  $\|\mathcal{M}(\cdot)\|_{\text{sp}}$ , between  $\Theta$  and its rank- $r$  projection,

$$f(r) = \inf\{\|\mathcal{M}(\Theta - \mathcal{A})\|_{\text{sp}} : \text{Rank}(\mathcal{A}) \leq (r, r, r)\}.$$

The tensor  $\Theta$  is called rank- $\sqrt{d}$  approximatable, if  $f(\sqrt{d}) \leq \sqrt{d}$ . Geometrically, the intersection point between two curves  $f(r)$  and  $g(r) = r$  is smaller than  $\sqrt{d}$ .

Equivalently,  $\Theta$  admits the decomposition

$$\Theta = \mathcal{A} + \mathcal{A}^\perp, \quad \text{s.t.} \quad \text{Rank}(\mathcal{A}) \leq (\sqrt{d}, \sqrt{d}, \sqrt{d}), \quad \text{and} \quad \|\text{Unfold}(\mathcal{A}^\perp)\|_{\text{sp}} \leq \sqrt{d}. \quad (1)$$

**Proposition 1** (Smooth matrix). Every Lipschitz smooth matrix is rank- $\sqrt{d}$  approximatable.

*Proof of Proposition 1.* Let  $\Theta$  be a Lipschitz smooth matrix. Set  $\mathcal{A} = \text{Block}(\Theta, \sqrt{d})$  and  $\mathcal{A}^\perp = \Theta - \text{Block}(\Theta, \sqrt{d})$ . Then, by approximation theorem,

$$\|\text{Unfold}(\mathcal{A}^\perp)\|_{\text{sp}} \leq \|\mathcal{A}^\perp\|_F \leq \sqrt{\frac{d^2}{d}} = \sqrt{d}.$$

Since  $\mathcal{A}$  is of rank at most  $\sqrt{d}$ , the decomposition satisfies the condition (1).  $\square$

**Conjecture 1** (Higher-order spectral algorithm). Suppose  $\Theta$  is an order-3, rank- $\sqrt{d}$  approximatable tensor. Then, the rank- $\sqrt{d}$  higher-order spectral algorithm [1] yields the estimate  $\hat{\Theta}$  with error bound

$$\mathcal{R}(\hat{\Theta}, \Theta) \lesssim d^{-1}.$$

Intuition: We decompose the error into estimation error and approximation bias

$$\begin{aligned} \|\hat{\Theta} - \Theta\|_F^2 &\leq \|\hat{\Theta} - \mathcal{A}\|_F^2 + \|\mathcal{A}^\perp\|_F^2 \\ &\lesssim \underbrace{(d^{3/2}r + dr^2 + r^3)}_{\text{by Proposition 1 in [1]}} + \underbrace{d[f(r)]^2}_{\leq d^2 \text{ by Assumption 1}} \\ &\lesssim d^2 \text{ if } r \asymp \sqrt{d}. \end{aligned}$$

More careful analysis is needed though, e.g. additive Gaussian vs. Bernoulli models, non-uniqueness of  $\mathcal{A}$  and its singular space, etc. Also, the rank choice  $\asymp \sqrt{d}$  is meaningful only in asymptotical sense. In practice, we should choose rank  $C\sqrt{d}$  where the constant  $C$  may depend on actual  $\Theta$ , noise, etc.

SBM (HOS+iteration)	sort-and-smoothing	square spectral	higher-order spectral (HOS)	NN
$d^{-6/5}$	$d^{-6/5}$ (restricted model)	$d^{-2/3}$	$d^{-1}$ (restricted model)	?

Table 1: Convergence rate for order-3 smooth tensors.

**Questions 1.** Unlike matrices, not every order-3 smooth tensor is  $\sqrt{d}$ -approximatable. How large is the order-3 tensor family that satisfy (1)? Does the signal tensor in our simulations satisfy (1)? How about general order- $m$  tensors? Fill in the rate for NN.

## 2 Block approximatable

Based on the proof of [1, Proposition 1], Conjecture 1 also applies to the block approximatable tensor. More generally, we aim to carve out the regimes for which HOS algorithm works.

**Definition 2** (Block- $d^\beta$  approximatable tensor). An order- $m$  tensor  $\Theta$  is called block approximatable with index  $\beta \in [0, 1]$  if it admits the decomposition  $\Theta = \mathcal{A} + \mathcal{A}^\perp$  satisfying the following two constraints:

1.  $\mathcal{A}$  is a  $d^\beta$ -block tensor;
2.  $\mathcal{A}^\perp$  has controlled spectral complexity in that

$$\|\mathcal{A}^\perp\|_{\text{sp}} \leq \sqrt{d}, \quad \text{and} \quad \|\text{Unfold}(\mathcal{A}^\perp)\|_{\text{sp}} \leq d^{\frac{m}{4}}. \quad (2)$$

By definition, every tensor is block approximatable with trivial  $\beta = 1$ . We make the convention that  $\beta$  denotes the minimal block complexity in the decomposition for which the residual tensor satisfy (2).

**Proposition 2** (Examples).

- Every Lipschitz smooth matrix is block approximatable with  $\beta = 1/2$ ;
- Every low-rank tensor with bounded factors has  $\beta = 0$  (conjecture).
- Gaussian random tensor has  $\beta \rightarrow 1$  for every  $m \geq 2$  (conjecture).

**Remark 2.** Not sure which of (2) and (1) has better intuitive interpretation. On one hand, the block assumption on  $\mathcal{A}$  is more restricted than the rank assumption. On the other hand, the spectral constraint on  $\mathcal{A}^\perp$  in (2) is more relaxed than (1), because  $\|\mathcal{A}^\perp\|_{\text{sp}} \leq \|\text{Unfold}(\mathcal{A}^\perp)\|_{\text{sp}}$  [2]. In both cases, we need the  $\|\text{Unfold}(\mathcal{A}^\perp)\|_{\text{sp}} \leq d^{m/4}$  for convergence guarantee of HOS algorithm [1, Proposition 1].

**Conjecture 2.** Suppose  $\Theta$  is a block approximatable tensor with  $\beta \leq \frac{m}{m+2}$ . Then the HOS algorithm in [1] with rank specification

$$r_* = \begin{cases} d^{\frac{1}{3}}, & \text{when } m = 2; \\ d^{\frac{1}{2}}, & \text{when } m = 3; \\ d^{\frac{m}{m+2}}, & \text{when } m \geq 4, \end{cases} \quad (3)$$

gives the estimator  $\hat{\Theta}$  with error rate

$$\begin{aligned} \mathcal{R}(\Theta, \hat{\Theta}) &\leq d^{-m} \{d^{\frac{m}{2}+\beta} + d^{\beta m} + \min(d^{m-2\beta}, d^{\frac{m}{2}+1})\} \\ &\leq \begin{cases} d^{-\frac{2}{3}}, & \text{when } m = 2; \\ d^{-1}, & \text{when } m = 3; \\ d^{-\frac{2m}{m+2}}, & \text{when } m \geq 4. \end{cases} \end{aligned}$$

**Questions 2.** What is the rate when  $\beta \geq \frac{m}{m+2}$ ? Implication in the matrix case. Compare with other methods in theory and in simulation. How large is the order-3 tensor family that satisfy (2)? Give two examples of smooth tensors that satisfy and violate this constraint, respectively. How about non-smooth tensors, e.g., single index tensors, glm tensors?

## 3 Intuition

- Oracle risk:

$$\underbrace{r^m}_{\text{block mean}} + \underbrace{d \log r}_{\text{block position}} \asymp \underbrace{\frac{d^m}{r^2}}_{m\text{-way approximation}}$$

Therefore, the best  $r \asymp d^{\frac{m}{m+2}}$ . When  $m = 2$  (matrix),  $r = \sqrt{d}$ .

- Oracle Spectral risk:

$$\underbrace{dr + r^m}_{\text{d.f. in spectral method}} \asymp \underbrace{\frac{d^m}{r^2}}_{m\text{-way approximation}}.$$

When  $m = 2$ , the left hand side is computable by matrix SVD. The best  $r = d^{-1/3}$ .

When  $m \geq 3$ , no polynomial time algorithm is able to solve exact SVD. The best-so-far polynomial algorithm increases the risk to

$$\underbrace{d^{m/2}r + r^m}_{\text{d.f. in spectral method}} \asymp \underbrace{\frac{d^m}{r^2}}_{m\text{-way approximation}}.$$

Notice the extra cost one has to pay on  $d$  when  $m \geq 3$ . The best  $r$  is solved in (3).

- NN risk for  $m = 2$ :

$$\underbrace{dr}_{\text{d.f. in row-based NN}} \asymp \underbrace{\frac{d^2}{r}}_{\text{row-based approximation}}.$$

The best  $r = \sqrt{d}$ , which yields the risk  $d^{-1/2}$ . Why  $\frac{d^2}{r}$  on right hand side? Because row-based NN partitions the rows into  $r$  groups, but the keep the  $d$  columns as they are. The accuracy is suboptimal even when the true two-way clustering patten is known a prior (check...).

## References

- [1] Rungang Han, Yuetian Luo, Miaoyan Wang, and Anru R Zhang, *Exact clustering in tensor block model: Statistical optimality and computational limit*, arXiv preprint arXiv:2012.09996 (2020).
- [2] Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S Song, *Operator norm inequalities between tensor unfoldings on the partition lattice*, Linear Algebra and Its Applications **520** (2017), 44–66.