The Optimization Landscape of Tensor Decompositions[1]

Presentation by Kiran Vodrahalli for ELE 538B
May 1, 2017

[1]Based on the work of Rong Ge and Tengyu Ma

Tensors: A brief definition

- Tensors are arrays indexed by multiple indices
- Each index represents a factor of interest
- Ex: Consider Netflix data over time
    - viewer
    - movie
    - time

Tensor Decomposition

- Goal: Find decomposition

$$T = \sum_{i=1}^{r} \lambda_i \vec{x}_i \otimes \vec{y}_i \otimes \vec{z}_i$$

- Optimization as iterative procedure
  - Find each component one-by-one

- Methods like gradient ascent and tensor power method work empirically well

Applications of Tensor Decomposition

- Latent variable models
  - HMMs
  - Gaussian mixture models
  - Topic modeling
  - ICA
  - and more…
- Symmetric Orthogonal Tensor Decomposition suffices for these models

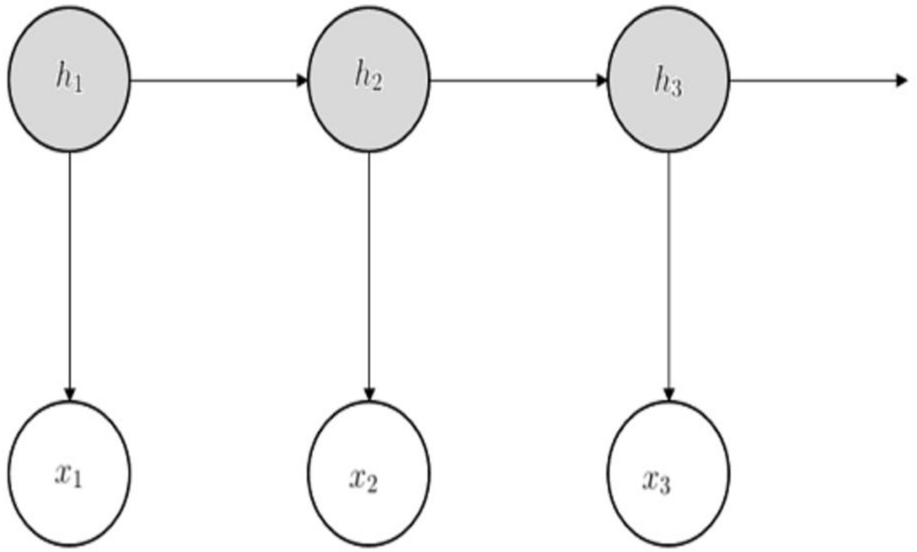
Hidden Markov Model

Ex: Trigrams in language modeling

Condition on middle topic l

x, y, z: conditional probabilities given topic l for each trigram position

$$T_{i,j,k} = \Pr[x_1 = i, x_2 = j, x_3 = k]$$



$$T = \sum_{l=1}^{n} \Pr[h_2 = l]\vec{x}_l \otimes \vec{y}_l \otimes \vec{z}_l.$$

Learning decompositions in the general case

- Sometimes, a true decomposition does not even exist
- Tensor problems tend to be NP-hard
- Motivates considering “average case” situations
    - $N \leq d$ and orthogonal components possible
    - What about $N >> d$ and non-orthogonal?

Provably learning **overcomplete** decompositions

$$\max \quad f(x) = \sum_{i,j,k,l \in [d]^4} T_{i,j,k,l} x_i x_j x_k x_l = \sum_{i=1}^{n} \langle a_i, x \rangle^4$$

$$s.t. \quad \|x\| = 1.$$

(multi-linear form)

under constraints:

- $a_i \in \mathbb{R}^d$ drawn i.i.d. from $\mathcal{N}(0, I)$
- $n \gg d$

Main Theorem

Theorem 1.1. *Let* $\varepsilon, \zeta \in (0, 1/3)$ *be two arbitrary constants and* d *be sufficiently large. Suppose* $d^{1+\varepsilon} < n < d^{2-\varepsilon}$. *Then, with high probability over the randomness of* a_i*'s, we have that in the superlevel set*

$$L = \left\{ x \in S^{d-1} : f(x) \geq 3(1+\zeta)n \right\}, \tag{1.2}$$

there are exactly $2n$ *local maxima with function values* $(1 \pm o(1))d^2$*, each of which is close to one of* $\pm\frac{1}{\sqrt{d}}a_1, \ldots, \pm\frac{1}{\sqrt{d}}a_n$.

- Initialization must be slightly better than random (function value 3n)
- Gradient ascent / power method then works
  - "Peel off eigenvectors" (c.f. SVD)

Proof Strategy

- Kac-Rice formula:
  - Assign probability to points on unit sphere of being local optima
  - Integrate to get expected # of optima
  - Need to analyze joint distribution of gradient and Hessian for local optimality
- Intractable closed form
- Estimate # local optima for:
  - "Local set": points near approximate optima
  - "Global set": everything else

Local-Global Set Decomposition

$$L_1 := \left\{ x \in S^{d-1} : \sum_{i=1}^{n} \langle a_i, x \rangle^4 \geq 3n + \gamma\sqrt{nd} \right\}$$

$$L_1 = (L_1 \cap L_2) \cup L_2^c,$$

$$\text{where } L_2 := \{x \in S^{d-1} : \forall i, \|P_x a_i\|^2 \geq (1-\delta)d, \quad \text{and} \quad |\langle a_i, x \rangle|^2 \leq \delta d\}$$

P_x is $(I - xx^T)$, the orthogonal projection operator.

Local Analysis ($L2^C$) uses RIP

- Local set is where both restricted isometry and approximate optimality hold
  - Intuitively, Gaussian components are "almost orthogonal" due to rotational invariance ⇒ RIP
  - Thus x has high correlation with only few components
- 2n local optima (+/- components)

- In high-correlation regions, objective is strongly convex with unique optimum

Global Analysis (L1 ∩ L2)

- Number of local optima is an integer r.v.

- If expected #optima << 1, Markov's inequality
  $\Rightarrow$ # optima is exactly 0 in this region w.h.p.

- Use random matrix theory on Kac-Rice integral to show required expectation result
  - analyze gradient and Hessian
  - Crux is determinant of Hessian analysis

Citations

- On the Optimization Landscape of Tensor Decompositions (Ge and Ma 2016)
- Tensor Decompositions for Learning Latent Variable Models (Anandkumar, Ge, Hsu, Kakade, Telgarsky)
- Tensor Methods in Machine Learning (Rong Ge, http://www.offconvex.org/2015/12/17/tensor-decompositions/)