Tensor power method

Daniel Hsu

COMS 4772

Orthogonal tensor decompositions

Moments of hidden variable models

- Many hidden variable models have observable moments (perhaps after transformation) of the form

$$\sum_{i=1}^{k} w_i \cdot \boldsymbol{\mu}_i^{\otimes p} .$$

- Jennrich's algorithm: uses $\boldsymbol{S}$ ($p = 2$) and $\boldsymbol{T}$ ($p = 3$) to recover parameters (assuming $\mathbb{S} := \operatorname{span}\{\boldsymbol{\mu}_i\}_{i=1}^{k}$ has dimension k).

Orthogonality

- Suppose all $w_i > 0$ and $\dim(\mathbb{S}) = k$.
- Then $\boldsymbol{S}$ is psd and has rank k.
- $\boldsymbol{S}$ defines inner product over $\mathbb{S}$ in which $\{\sqrt{w_i}\boldsymbol{\mu}_i\}_{i=1}^{k}$ are *orthonormal*:

$$\begin{aligned} \langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{S}^\dagger} &:= \boldsymbol{x}^\top \boldsymbol{S}^\dagger \boldsymbol{y} . \\ \langle \sqrt{w_i}\boldsymbol{\mu}_i, \sqrt{w_j}\boldsymbol{\mu}_j \rangle_{\boldsymbol{S}^\dagger} &= \begin{cases} 1 & \text{if } i = j , \\ 0 & \text{if } i \neq j . \end{cases} \end{aligned}$$

Whitening

- Can write $\boldsymbol{S}^{\dagger} := \boldsymbol{W}\boldsymbol{W}^{\top}$ with rank k matrix $\boldsymbol{W} \in \mathbb{R}^{d \times k}$ called "whitening transformation":

$$\boldsymbol{S}(\boldsymbol{W}, \boldsymbol{W}) \;=\; \boldsymbol{W}^{\top}\boldsymbol{S}\boldsymbol{W} \;=\; \boldsymbol{I}\,,$$

  so $\{\boldsymbol{W}^{\top}(\sqrt{w_i}\boldsymbol{\mu}_i)\}_{i=1}^{k}$ is ONB in $\mathbb{R}^k$.
- Can also apply $\boldsymbol{W}$ to higher-order tensors, e.g.,

$$\begin{aligned}\boldsymbol{T}(\boldsymbol{W}, \boldsymbol{W}, \boldsymbol{W}) \;&=\; \sum_{i=1}^{k} w_i \cdot (\boldsymbol{W}^{\top}\boldsymbol{\mu}_i)^{\otimes 3} \\ &=\; \sum_{i=1}^{k} \frac{1}{\sqrt{w_i}} \cdot (\boldsymbol{W}^{\top}(\sqrt{w_i}\boldsymbol{\mu}_i))^{\otimes 3}\,.\end{aligned}$$

Odeco tensors

- (Symmetric) *orthogonally decomposable (odeco) tensors*:

$$\sum_{i=1}^{k} \lambda_i \boldsymbol{v}_i^{\otimes p}$$

  where $\lambda_i > 0$ and $\{\boldsymbol{v}_i\}_{i=1}^{k}$ is ONB.
  - (Assume positivity of λ_i for simplicity.)
- Is the decomposition of an odeco tensor unique?
  - $p = 2$: **no**
  - $p \geq 3$: **yes**
- **Variational claim**: for $p \geq 3$, isolated local maximizers of degree-p homogeneous polynomial $f_{\boldsymbol{T}}(\boldsymbol{x}) := \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \ldots, \boldsymbol{x})$ over B^k are $\{\boldsymbol{v}_i\}_{i=1}^{k}$.

Variational characterization

- **Claim**: for $p \geq 3$, isolated local maximizers of $f_{\boldsymbol{T}}(\boldsymbol{x}) := \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \ldots, \boldsymbol{x})$ over B^k are $\{\boldsymbol{v}_i\}_{i=1}^k$.
- Observation: by orthogonality,

$$f_{\boldsymbol{T}}(\boldsymbol{v}_j) \;=\; \sum_{i=1}^{k} \lambda_i \langle \boldsymbol{v}_j, \boldsymbol{v}_i \rangle^p \;=\; \lambda_j\,.$$

  - What about other vectors?
- May as well think of $\boldsymbol{v}_i$ as i-th coordinate basis vector.

$$\max_{\boldsymbol{x} \in \mathbb{R}^k} \sum_{i=1}^{k} \lambda_i x_i^p \quad \text{s.t.} \quad \sum_{i=1}^{k} x_i^2 \;\leq\; 1\,.$$

- If both x_1 and x_2 are non-zero, then

$$\lambda_1 x_1^p + \lambda_2 x_2^p \;<\; \lambda_1 x_1^2 + \lambda_2 x_2^2 \;\leq\; \max\{\lambda_1, \lambda_2\}\,.$$

  - Hence, better to only have a *single* non-zero entry.
  - I.e., better to have $\boldsymbol{x} = \boldsymbol{v}_i$ for some i. $\square$

Tensor power method

Optimality condition

$$\max_{\boldsymbol{x} \in \mathbb{R}^k} \sum_{i=1}^{k} \lambda_i \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle^p \quad \text{s.t.} \quad \sum_{i=1}^{k} x_i^2 \;\leq\; 1 \,.$$

- Lagrangian:

$$\mathcal{L}(\boldsymbol{x}, \lambda) \;:=\; \sum_{i=1}^{k} \lambda_i \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle^p - \frac{p}{2} \lambda (\|\boldsymbol{x}\|_2^2 - 1) \,.$$

- First-order optimality condition:

$$p \sum_{i=1}^{k} \lambda_i \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle^{p-1} \boldsymbol{v}_i - p\lambda \boldsymbol{x} \;=\; \boldsymbol{0} \,.$$

  - I.e.,

$$\boldsymbol{T}(\underbrace{\boldsymbol{x}, \dots, \boldsymbol{x}}_{p-1 \text{ times}}, \boldsymbol{I}) \;=\; \sum_{i=1}^{k} \lambda_i \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle^{p-1} \boldsymbol{v}_i \;=\; \lambda \boldsymbol{x} \,.$$

  - Maximizer must be an "eigenvector" of degree-($p-1$) map.

Fixed-point iteration algorithm

- Consider map from first-order condition:

$$\phi_{\boldsymbol{T}}(\boldsymbol{x}) \;:=\; \boldsymbol{T}(\underbrace{\boldsymbol{x}, \dots, \boldsymbol{x}}_{p-1 \text{ times}}, \boldsymbol{I}) \,.$$

- Goal: find $\boldsymbol{x} \in S^{k-1}$ that is fixed under

$$\boldsymbol{x} \;\mapsto\; \frac{\phi_{\boldsymbol{T}}(\boldsymbol{x})}{\|\phi_{\boldsymbol{T}}(\boldsymbol{x})\|_2} \,.$$

- "Tensor power method" (De Lathauwer et al, 2000):
  - Repeatedly apply $\phi_{\boldsymbol{T}}$ to initial $\boldsymbol{x}^{(0)} \in S^{k-1}$ (and re-normalize).
- Question: Does it find the $\boldsymbol{v}_i$?

Example

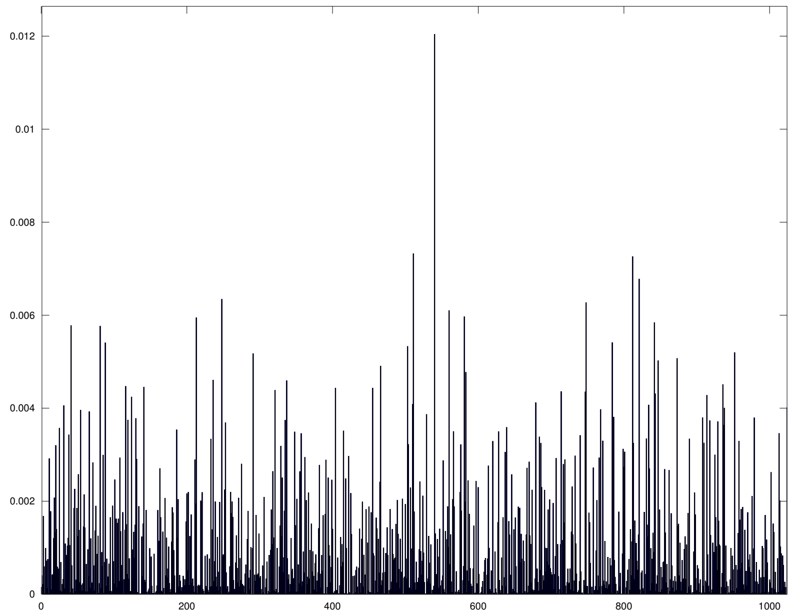

Figure 1: $\langle \boldsymbol{x}^{(0)}, \boldsymbol{v}_i \rangle$ for $i = 1, 2, \ldots, 1024$

Example

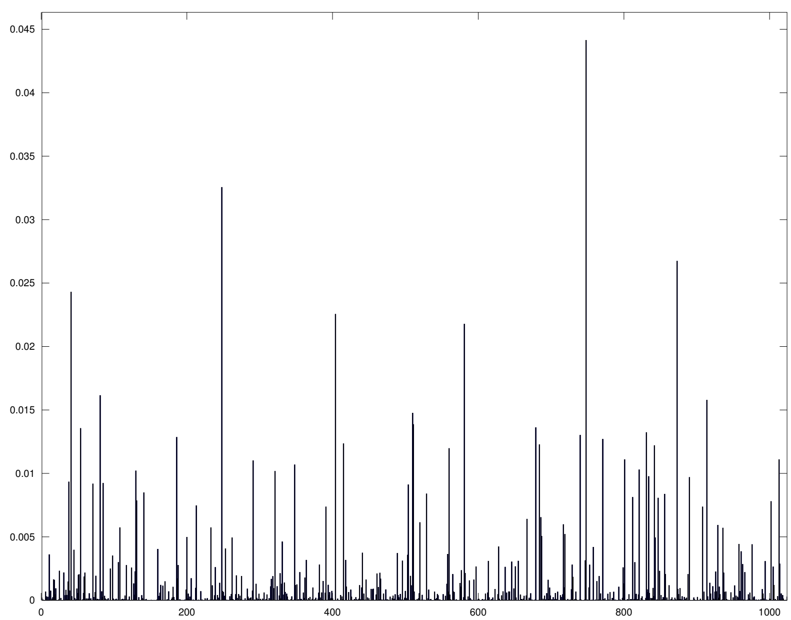

Figure 2: $\langle \boldsymbol{x}^{(1)}, \boldsymbol{v}_i \rangle$ for $i = 1, 2, \ldots, 1024$

Example

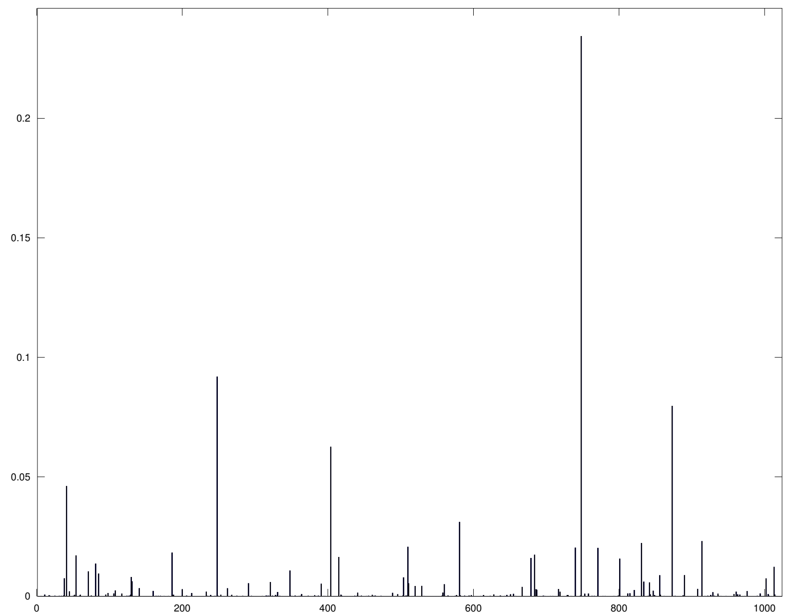

Figure 3: $\langle \boldsymbol{x}^{(2)}, \boldsymbol{v}_i \rangle$ for $i = 1, 2, \ldots, 1024$

Example

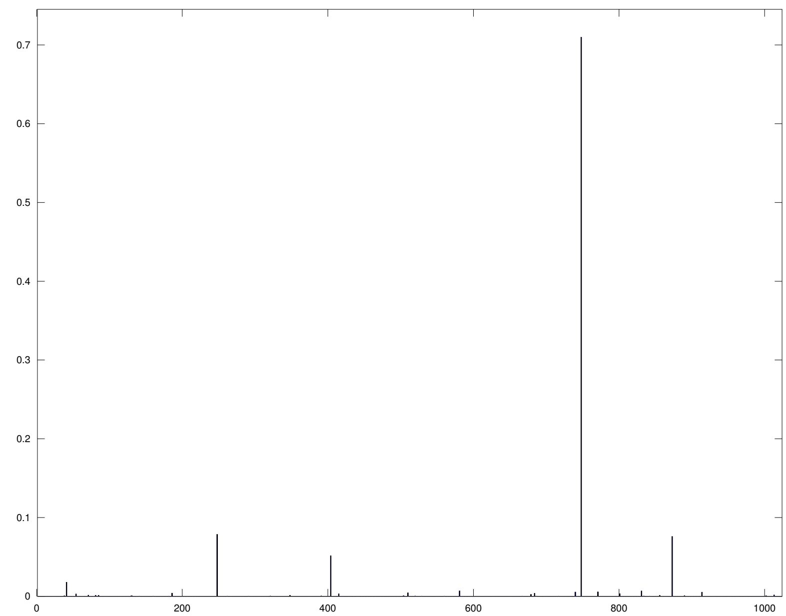

Figure 4: $\langle \boldsymbol{x}^{(3)}, \boldsymbol{v}_i \rangle$ for $i = 1, 2, \ldots, 1024$

Example

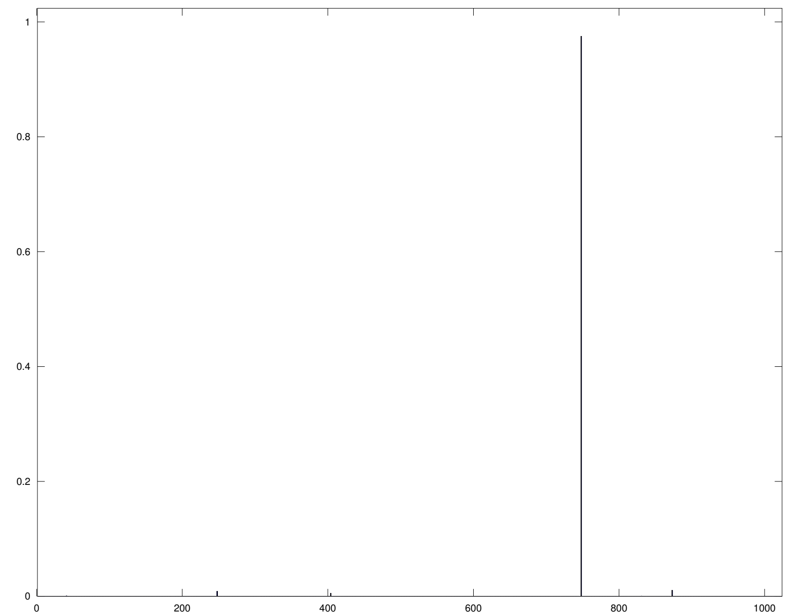

Figure 5: $\langle \boldsymbol{x}^{(4)}, \boldsymbol{v}_i \rangle$ for $i = 1, 2, \ldots, 1024$

Example

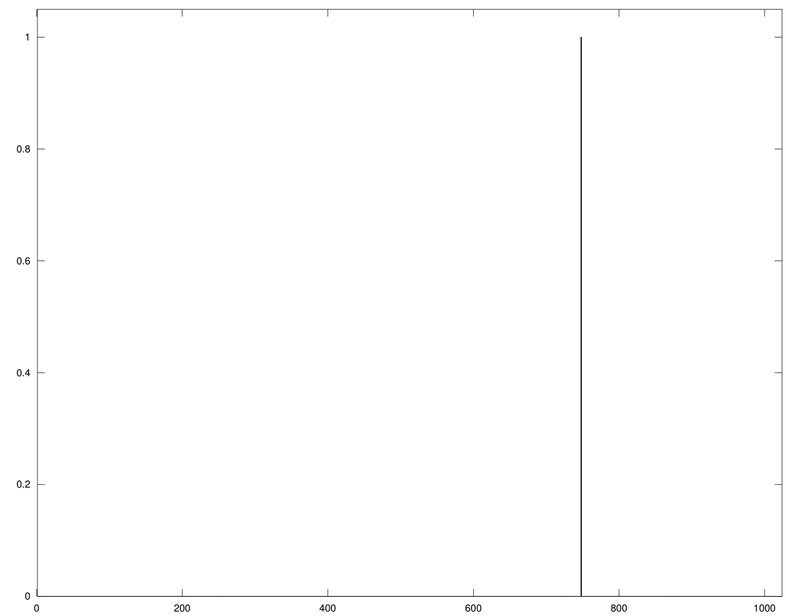

Figure 6: $\langle \boldsymbol{x}^{(5)}, \boldsymbol{v}_i \rangle$ for $i = 1, 2, \ldots, 1024$

Review: matrix power method

- Tensor power method for $p = 2$ is "matrix power method":

$$\boldsymbol{x}^{(t)} \ := \ \boldsymbol{M}(\boldsymbol{I}, \boldsymbol{x}^{(t-1)}) \ = \ \boldsymbol{M}\boldsymbol{x}^{(t-1)} .$$

- If $\lambda_1 > \lambda_2$, and $\boldsymbol{x}^{(0)}$ not orthogonal to $\boldsymbol{v}_1$, then angle $\theta^{(t)}$ between $\boldsymbol{x}^{(t)}$ and $\boldsymbol{v}_1$ decreases to zero at linear rate.
  - Write $c_i := \langle \boldsymbol{x}^{(0)}, \boldsymbol{v}_i \rangle$ for $i = 1, 2, \dots, k$, so

$$\begin{aligned} \boldsymbol{x}^{(0)} &= c_1 \boldsymbol{v}_1 + c_2 \boldsymbol{v}_2 + \cdots + c_k \boldsymbol{v}_k , \\ \boldsymbol{x}^{(t)} &= \lambda_1^t c_1 \boldsymbol{v}_1 + \lambda_2^t c_2 \boldsymbol{v}_2 + \cdots + \lambda_k^t c_k \boldsymbol{v}_k . \end{aligned}$$

  - $\cos^2(\theta^{(t)})$:

$$\frac{\langle \boldsymbol{x}^{(t)}, \boldsymbol{v}_1 \rangle^2}{\|\boldsymbol{x}^{(t)}\|_2^2} \ = \ \frac{c_1^2 \lambda_1^{2t}}{\sum_{i=1}^k (c_i \lambda_i^t)^2} \ \geq \ \frac{1}{1 + \left(\frac{\lambda_2}{\lambda_1}\right)^{2t} \frac{1-c_1^2}{c_1^2}} .$$

- $p = 2$ behavior very different from $p \geq 3$.

Tensor power method ($p = 3$)

- Re-number components so that

$$\lambda_1 |c_1| \ \geq \ \lambda_2 |c_2| \ \geq \ \cdots .$$

- Then

$$\boldsymbol{x}^{(1)} \ = \ \sum_{i=1}^{k} \lambda_i \langle \boldsymbol{x}^{(0)}, \boldsymbol{v}_i \rangle^2 \boldsymbol{v}_i \ = \ \sum_{i=1}^{k} \lambda_i c_i^2 \boldsymbol{v}_i .$$

  - Coefficient c_i in $\boldsymbol{x}^{(0)}$ is *squared* in $\boldsymbol{x}^{(1)}$.
- If $\lambda_1 |c_1| > \lambda_2 |c_2|$, then angle between $\boldsymbol{x}^{(t)}$ and $\boldsymbol{v}_1$ decreases to zero at *quadratic rate*:

$$\frac{\langle \boldsymbol{x}^{(t)}, \boldsymbol{v}_1 \rangle^2}{\|\boldsymbol{x}^{(t)}\|_2^2} \ \geq \ \frac{1}{1 + \left(\frac{\lambda_2 |c_2|}{\lambda_1 |c_1|}\right)^{2^{t+1}} \sum_{i=2}^k \left(\frac{\lambda_1}{\lambda_i}\right)^2} .$$

- Note: which vector we called $\boldsymbol{v}_1$ depends on $\boldsymbol{x}^{(0)}$!

Initialization of tensor power method

- Convergence of tensor power method requires gap between largest and second-largest $\lambda_i |\langle \boldsymbol{x}^{(0)}, \boldsymbol{v}_i \rangle|^{p-2}$.
- Bad initialization:
  - Suppose $\boldsymbol{T} = \sum_{i=1}^{k} \boldsymbol{v}_i^{\otimes p}$ and $\boldsymbol{x}^{(0)} = \boldsymbol{v}_1 + \boldsymbol{v}_2$:

$$\begin{aligned} \phi_{\boldsymbol{T}}(\boldsymbol{x}^{(0)}) &= \langle \boldsymbol{x}^{(0)}, \boldsymbol{v}_1 \rangle^p \boldsymbol{v}_1 + \langle \boldsymbol{x}^{(0)}, \boldsymbol{v}_2 \rangle^p \boldsymbol{v}_2 \\ &= \boldsymbol{v}_1 + \boldsymbol{v}_2 \,. \end{aligned}$$

- But bad initialization points comprise measure-zero set.

Recovering all components

- Power method with $\boldsymbol{T} = \sum_{i=1}^{k} \lambda_i \boldsymbol{v}_i^{\otimes p}$ returns some $\boldsymbol{v}_i$.
  - Can also get λ_i via $\lambda_i = \boldsymbol{T}(\boldsymbol{v}_i, \boldsymbol{v}_i, \dots, \boldsymbol{v}_i)$.
- What about other components $j \neq i$?
  - "Deflation": replace $\boldsymbol{T}$ with $\boldsymbol{T}' := \boldsymbol{T} - \lambda_i \boldsymbol{v}_i^{\otimes p}$ so that

$$\boldsymbol{T}' = \sum_{j \neq i} \lambda_j \boldsymbol{v}_i^{\otimes p} \,.$$

  - Can do this "inside" power method:

$$\boldsymbol{T}'(\boldsymbol{x}, \dots, \boldsymbol{x}, \boldsymbol{I}) = \boldsymbol{T}(\boldsymbol{x}, \dots, \boldsymbol{x}, \boldsymbol{I}) - \lambda_i \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle^{p-1} \boldsymbol{v}_i \,.$$

  - Implicitly tries to make power method (with $\boldsymbol{T}'$) converge to something orthogonal to $\boldsymbol{v}_i$.
  - Caveat: don't have $\boldsymbol{v}_i$ and λ_i *exactly*, but only up to some small error, e.g.,

$$\|\hat{\boldsymbol{v}}_i - \boldsymbol{v}_i\|_2 \leq \varepsilon \,, \quad |\hat{\lambda}_i - \lambda_i| \leq \varepsilon' \,.$$

Error analysis

Nearly odeco tensors

- Suppose we have $\widehat{\boldsymbol{T}} = \boldsymbol{T} + \boldsymbol{E}$ for some odeco $\boldsymbol{T} = \sum_{i=1}^{k} \lambda_i \boldsymbol{v}_i^{\otimes p}$ and (symmetric) "error tensor" $\boldsymbol{E}$ with $\|\boldsymbol{E}\|_2 \leq \epsilon$, i.e.,

$$\max_{\boldsymbol{u} \in S^{k-1}} |\boldsymbol{E}(\boldsymbol{u}, \boldsymbol{u}, \ldots, \boldsymbol{u})| \ \leq \ \epsilon \,.$$

- Matrix case ($p = 2$): ($\lambda_1 \geq \lambda_2 \geq \cdots$)
  - Top eigenvalue/eigenvector $(\hat{\lambda}, \hat{\boldsymbol{v}})$ of $\widehat{\boldsymbol{T}}$.
  - $\hat{\lambda}$ approximates λ_1:
  $$|\hat{\lambda} - \lambda_1| \ \leq \ \epsilon \,.$$
  - But need $\epsilon < \lambda_1 - \lambda_2$ for $\hat{\boldsymbol{v}}$ to approximate $\boldsymbol{v}_1$ (Davis-Kahan).

Nearly odeco tensors ($p \geq 3$)

- Higher-order case ($p \geq 3$):
  - Maximum of $f_{\widehat{\boldsymbol{T}}}$ approximates some λ_i, i.e.,

$$\left| \max_{\boldsymbol{u} \in S^{k-1}} \widehat{\boldsymbol{T}}(\boldsymbol{u}, \boldsymbol{u}, \ldots, \boldsymbol{u}) - \lambda_i \right| \leq \epsilon .$$

  - Maximizers $\hat{\boldsymbol{v}}$ of $f_{\widehat{\boldsymbol{T}}}$ also approximate some λ_i, i.e.,

$$\|\hat{\boldsymbol{v}} - \boldsymbol{v}_i\|_2 \leq O\left(\frac{\epsilon}{\lambda_i} + \left(\frac{\epsilon}{\lambda_i} \right)^2 \right) .$$

  - Output of power method: depends on initialization

$$\boldsymbol{x}^{(0)} = c_1 \boldsymbol{v}_1 + c_2 \boldsymbol{v}_2 + \cdots + c_k \boldsymbol{v}_k .$$

  - E.g., if all $\lambda_i \in \left[\Omega(1), O(1)\right]$, then need $\max_i c_i^2 \gg \epsilon$ to get

$$|\hat{\lambda} - \lambda_i| \leq O(\epsilon) , \qquad \|\hat{\boldsymbol{v}} - \boldsymbol{v}_i\|_2 \leq O(\epsilon)$$

    for some component i, after $O(\log(k) + \log\log(1/\epsilon))$ iterations.

Error from deflation

- Since $(\hat{\lambda}, \hat{\boldsymbol{v}})$ obtained from $\widehat{\boldsymbol{T}} = \boldsymbol{T} + \boldsymbol{E}$ is not exactly $(\lambda_i, \boldsymbol{v}_i)$ for any component i of $\boldsymbol{T}$, "deflation" introduces some error:

$$\begin{aligned} \widehat{\boldsymbol{T}}' &:= \widehat{\boldsymbol{T}} - \hat{\lambda} \hat{\boldsymbol{v}}^{\otimes p} \\ &= \sum_{j=1}^{k} \lambda_j \boldsymbol{v}_j^{\otimes p} + \boldsymbol{E} - \hat{\lambda} \hat{\boldsymbol{v}}^{\otimes p} \\ &= \sum_{j \neq i} \lambda_j \boldsymbol{v}_j^{\otimes p} + \boldsymbol{E} + \left(\lambda_i \boldsymbol{v}_i^{\otimes p} - \hat{\lambda} \hat{\boldsymbol{v}}^{\otimes p} \right) \\ &=: \boldsymbol{T}' + \boldsymbol{E} + \boldsymbol{E}_i . \end{aligned}$$

- **Danger**: $\|\boldsymbol{E}_i\|_2$ can be as large as $\|\boldsymbol{E}\|_2$.
  - So "error" has doubled?

Analysis of deflation error

- (For simplicity, assume all $\lambda_i = 1$.)
- Deflation error: $\boldsymbol{E}_i = \boldsymbol{v}_i^{\otimes p} - \hat{\boldsymbol{v}}^{\otimes p}$
  - All we know is that $\|\hat{\boldsymbol{v}} - \boldsymbol{v}_i\|_2 \leq O(\epsilon)$.
- Consider a unit vector $\boldsymbol{u}$ orthogonal to $\boldsymbol{v}_i$:

$$\begin{aligned} \|\phi_{\boldsymbol{E}_i}(\boldsymbol{u})\|_2 &= \|\langle \boldsymbol{u}, \boldsymbol{v}_i\rangle^{p-1}\boldsymbol{v}_1 - \langle \boldsymbol{u}, \hat{\boldsymbol{v}}\rangle^{p-1}\hat{\boldsymbol{v}}\|_2 \\ &= \|\langle \boldsymbol{u}, \hat{\boldsymbol{v}}\rangle^{p-1}\hat{\boldsymbol{v}}\|_2 \\ &= |\langle \boldsymbol{u}, \hat{\boldsymbol{v}} - \boldsymbol{v}_i\rangle|^{p-1} \\ &\leq O(\epsilon^{p-1}). \end{aligned}$$

- Therefore, for such $\boldsymbol{u}$,

$$\|\phi_{\boldsymbol{E}+\boldsymbol{E}_i}(\boldsymbol{u})\|_2 \leq (1 + O(\epsilon^{p-2}))\epsilon.$$

- When $p \geq 3$, errors due to deflation have **lower-order effect** on ability to approximate remaining components.
  - Not true for $p = 2$.

Recap

- High-order moments
  - Get parameter identifiability.
  - But for very high-order moments, estimation may be difficult.
- Higher-than-order-2 moments in high-dimensions
  - Can still get parameter identifiability in many cases.
  - Arrangement in higher-order tensor facilitates reasoning/computation.
- Higher-than-order-2 tensors
  - Most computational problems (that were easy for matrices) become hard.
  - But when there is a lot of structure, some computational issues are better than in matrix case!