SAND2017-7716D

Exceptional service in the national interest

Sandia National Laboratories

Generalized Tensor Decompositions

David Hong[1,2] Mentors: Cliff Anderson-Bergman[1] and Tamara G. Kolda[1]

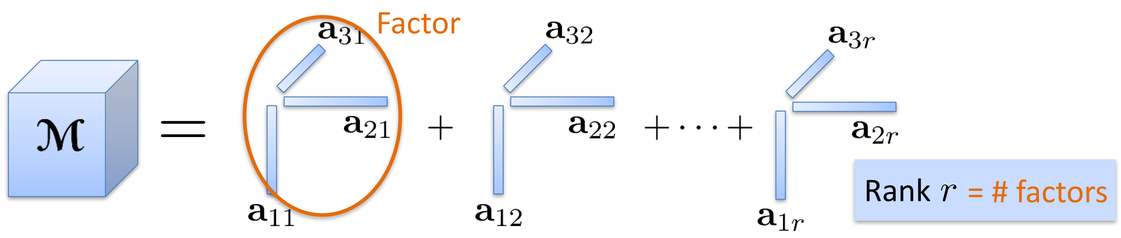
1 Introduction/Overview

Given a tensor, a CP decomposition is formed by finding a low-rank tensor approximation that has good fit as measured by the total entry-wise square difference. However, different fit/loss functions may be more appropriate in some cases. For example, logistic loss is a natural choice for binary tensors.

We propose a new generalized tensor decomposition method that allows users to select a generic loss function. To solve the resulting optimization problem, we use a stochastic gradient algorithm from machine learning to exploit the fact that approximate gradients can be computed efficiently from small samples of the entries.

2 Proposed Method: Generalized Tensor Decomposition

Find a low-rank tensor



$$\mathcal{M}_{ijk} = \mathbf{a}_{11}(i)\mathbf{a}_{21}(j)\mathbf{a}_{31}(k) + \mathbf{a}_{12}(i)\mathbf{a}_{22}(j)\mathbf{a}_{32}(k) + \cdots + \mathbf{a}_{1r}(i)\mathbf{a}_{2r}(j)\mathbf{a}_{3r}(k)$$

that minimizes

$$(\bigstar) \quad \min \quad F(\mathcal{X}, \mathcal{M}) = \sum_i f(x_i, m_i) \quad \text{s.t.} \quad \mathcal{M} = [\![\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_d]\!]$$

Shorthand

Different $f(x,m)$ capture different statistical assumptions:

$f(x,m) = (x-m)^2$ — "Standard" least squares/Gaussian
$\mathcal{X}_{ijk}$ is Gaussian with mean $\mathcal{M}_{ijk}$ and variance σ^2.

$f(x,m) = \log(1+e^m) - xm$ — Logistic/Bernoulli log-odds
$\mathcal{X}_{ijk}$ is Bernoulli with log-odds $\mathcal{M}_{ijk}$.

$f(x,m) = \log(m+1) - x\log m, m \geq 0$ — Bernoulli odds
$\mathcal{X}_{ijk}$ is Bernoulli with odds $\mathcal{M}_{ijk}$.

...many other options

Want: A fast and flexible algorithm to try all of the above.

3 Proposed Algorithm: Stochastic Gradient Methods

Solving (★) is challenging in general, but it turns out that sub-sampled gradients can be computed efficiently/cheaply!

Idea: Use stochastic gradient methods (Adam) from machine learning.

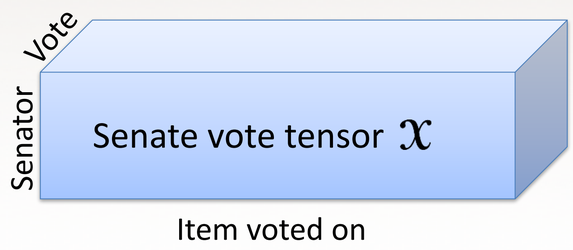
Given data tensor $\mathcal{X}$ and initial guess $\mathcal{M} = [\![\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_d]\!]$.

1: $\hat{\Omega} \leftarrow \hat{N}$ random entries in the data tensor
2: $F_0 \leftarrow \sum_{i \in \hat{\Omega}} f(x_i, m_i)$
3: $\Phi_k \leftarrow 0$, $\Psi_k \leftarrow 0$ for $k = 1, \ldots, d$
4: **for** $\ell = 1, 2, \ldots$ **do**
5: **for** $t = 1, \ldots, T$ **do**
6: $\Omega \leftarrow N$ random entries in the data tensor — Different samples every time
7: Compute $\mathcal{G}$ *only at entries in* Ω (everywhere else is zero)
8: $\Delta_k \leftarrow \mathbf{G}_{(k)}(\mathbf{A}_d \odot \cdots \odot \mathbf{A}_{k+1} \odot \mathbf{A}_{k-1} \odot \cdots \odot \mathbf{A}_1)$ for $k = 1, \ldots, d$ — Cheap for a small # of samples! Compute stochastic gradient
9: $\Phi_k \leftarrow \beta_1 \Phi_k + (1-\beta_1)\Delta_k$, $\hat{\Phi}_k \leftarrow \Phi_k / (1-\beta_1^{(\ell-1)T+t})$ for $k = 1, \ldots, d$
10: $\Psi_k \leftarrow \beta_2 \Psi_k + (1-\beta_2)\Delta_k^2$, $\hat{\Psi}_k \leftarrow \Psi_k / (1-\beta_2^{(\ell-1)T+t})$ for $k = 1, \ldots, d$ — Update moments
11: $\mathbf{A}_k \leftarrow \mathbf{A}_k - \alpha \hat{\Phi}_k / (\sqrt{\hat{\Psi}_k} + \varepsilon)$ for $k = 1, \ldots, d$ — Take a step
12: **end for**
13: $F_\ell \leftarrow \sum_{i \in \hat{\Omega}} f(x_i, m_i)$ using updated $\mathcal{M}$
14: Check convergence — Same samples every time
15: **end for**

Default parameter values
$\alpha = 10^{-3}$ $\beta_1 = 0.9$
$\varepsilon = 10^{-8}$ $\beta_2 = 0.999$

4 An illustrative example: Senate voting

Consider the following tensor (i.e., block) of data



The (i,j,k)th entry is
- 1 if senator i on item j casted vote k
- 0 otherwise

Question: Are there some latent factors that could explain the data? (e.g., political party)

Data from 1989-2016
271 Senators
- Two per state
- Six-year term (can serve multiple terms)
- Two main parties: Republicans and Democrats

9044 items voted on (roll call votes)

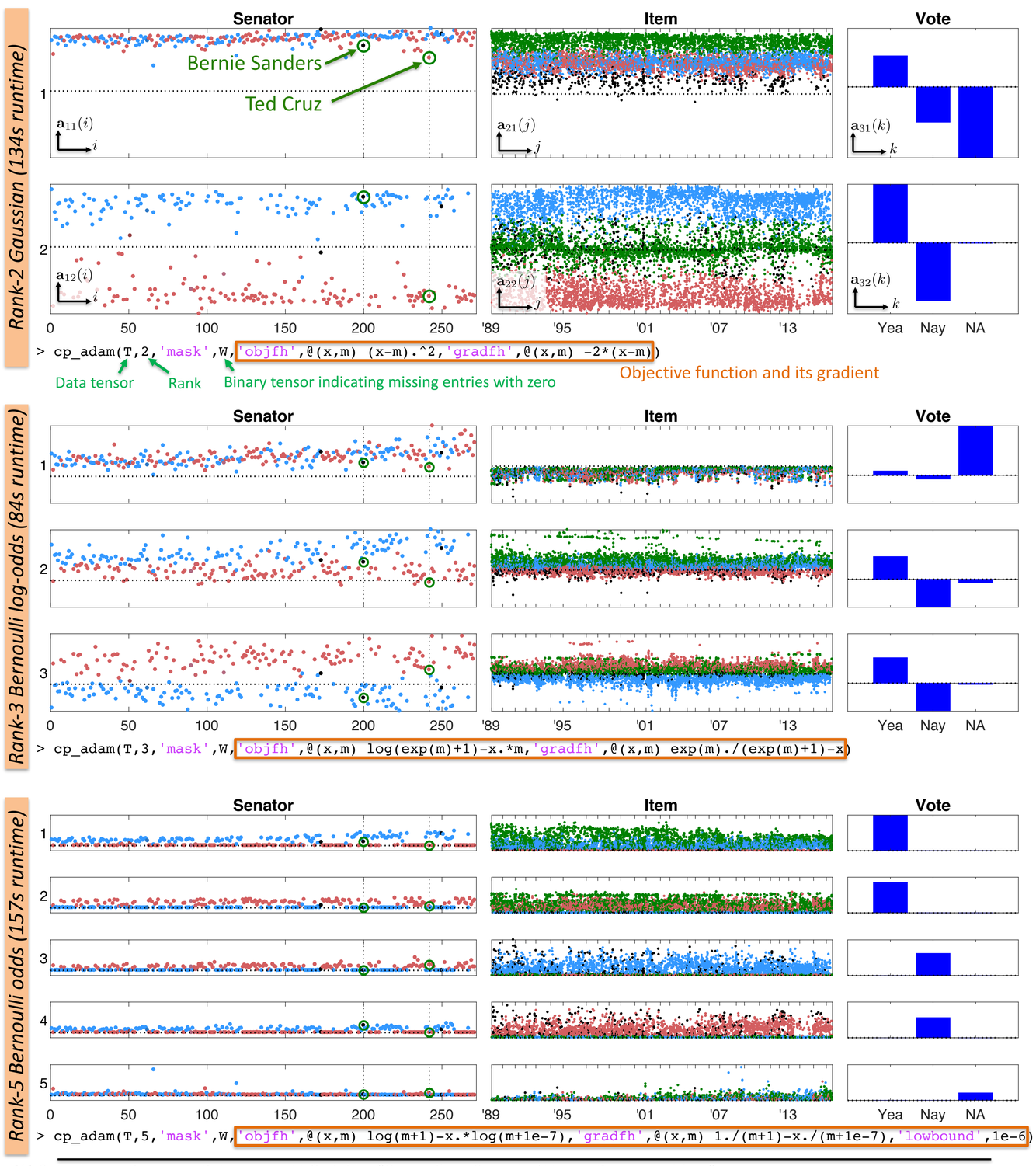
3 possible votes: yea, nay, no vote

63% entries not observed
(2.7 million observed / 7.4 million)

Senator Color
- Republican
- Democrat
- Independent

Item Color
- Supported by a majority of both parties
- Supported by a majority of Republicans only
- Supported by a majority of Democrats only
- Not supported by a majority of either party

Parameter selection
Defaults used for $\alpha, \varepsilon, \beta_1, \beta_2$
$\hat{N} = 2 \times 10^5$ (approx. 10% of observed entries)
$N = 10^4$ (approx. the sum of dimensions)
$T = 2 \times 10^3$ (chosen experimentally)



[1] C. Anderson-Bergman, D. Hong and T. G. Kolda. "Generalized Canonical Tensor Decomposition." In progress.
[2] T. G. Kolda and B. W. Bader. "Tensor decompositions and applications." *SIAM review* 51.3 (2009).
[3] E. C. Chi and T. G. Kolda. "On tensors, sparsity, and nonnegative factorizations." *SIAM Journal on Matrix Analysis and Applications* 33.4 (2012).
[4] M. Collins, S. Dasgupta and R. E. Schapire. "A generalization of principal components analysis to the exponential family." *Advances in neural information processing systems*. 2002.
[5] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization." *arXiv:1412.6980 (2014).*

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

[1]Sandia National Labs [2]University of Michigan, Department of EECS