CS 598 EVS: Tensor Computations

Tensor Decomposition

Edgar Solomonik

University of Illinois at Urbana-Champaign

CP Decomposition Rank

- The *canonical polyadic or CANDECOMP/PARAFAC (CP) decomposition* expresses an order d tensor in terms of d factor matrices

Tensor Rank Properties

- Tensor rank does not satisfy many of the properties of matrix rank

Typical Rank and Generic Rank

- When there is only a single typical tensor rank, it is the *generic rank*

Uniqueness Sufficient Conditions

- Unlike the low-rank matrix case, the CP decomposition can be unique

Uniqueness Necessary Conditions

- Necessary conditions for uniqueness of the CP decomposition also exist

Degeneracy

- The best rank-k approximation may not exist, a problem known as *degeneracy* of a tensor

Border Rank

- Degeneracy motivates an approximate notion of rank, namely *border rank*

Approximation by CP Decomposition

- Approximation via CP decomposition is a nonlinear optimization problem

Alternating Least Squares Algorithm

- The standard approach for finding an approximate or exact CP decomposition of a tensor is the *alternating least squares (ALS) algorithm*

Properties of Alternating Least Squares for CP

Alternating Least Squares for Tucker Decomposition

- For Tucker decomposition, an analogous optimization procedure to ALS is referred to as *high-order orthogonal iteration (HOOI)*

Dimension Trees for ALS

- The cost of ALS can be reduced by amortizing computation common terms

Fast Residual Norm Calculation

- Calculating the norm of the residual has cost $2ds^dR$, but can be done more cheaply within ALS

Pairwise Perturbation Algorithm

- A route to further reducing the cost of ALS is to perform it approximately via *pairwise perturbation*

Pairwise Perturbation Second Order Correction

- When approximating a tensor using CP, the partially converged CP factors can sometimes be used in place of the tensor to accelerate cost

Gauss-Newton Algorithm

- ALS generally achieves linear convergence, while Newton-based methods can converge quadratically

Gauss-Newton for CP Decomposition

- CP decomposition for order $d = 3$ tensors ($d > 3$ is similar) minimizes

Gauss-Newton for CP Decomposition

- A step of Gauss-Newton requires solving a linear system with $\boldsymbol{H}$

```
u = []
for q in range(d):
  u.append(zeros((n,R)))
  for p in range(d):
    if q == p:
      u[q] += einsum("rz,kz->kr",G[q,p],v[p])
    else:
      u[q] += einsum("kz,lr,rz,lz->kr", \
                     U[q],U[p],G[q,p],v[p])
```

Matrix Sketching

Randomized methods provide accurate approximate solutions to linear least squares problems, which can be applied to accelerate ALS, as well as more basic problems

Random Projections

Accuracy of sketching techniques is theoretically characterized by statistical analysis

Johnson-Lindenstrauss Lemma

The Johnson-Lindenstrauss lemma is a powerful tool for obtaining error bounds in a projected vector space

$$\boldsymbol{S}\boldsymbol{A}\hat{\boldsymbol{x}} \cong \boldsymbol{S}\boldsymbol{b}$$

Matrix Sketching

The best choice of sketch matrix depends on the desired accuracy and the structure of $\boldsymbol{A}$

Matrix Sketching via Sampling

Uniform sampling of rows is insufficient to obtain good accuracy in general

Leverage score sampling provides better accuracy guarantees

Mixing Techniques

To circumvent leverage score sampling, we can mix rows randomly Instead of

choosing elements of S randomly, pseudo-random distributions allow S to be applied more rapidly

Approximate CP ALS using Random Sampling

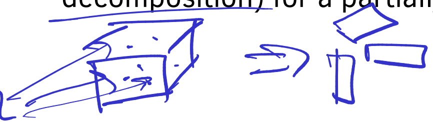
- Another approach to approximating ALS is to sample the least-squares equations[1]

[1] C. Battaglino, G. Ballard, T. G. Kolda, 2018

Tensor Completion

- The *tensor completion* problem seeks to build a model (e.g., CP or Tucker decomposition) for a partially-observed tensor

$$\sum_{r=1}^{R} u_{ir} v_{jr} w_{kr}$$



Ω

$\downarrow$

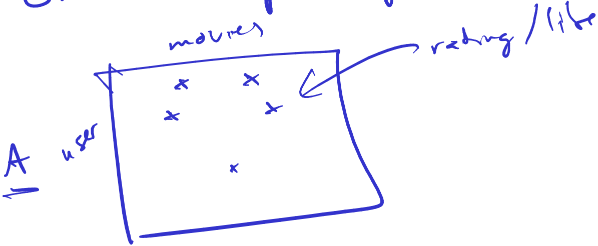
(i,j,k)

$$f(U,V,W) = \frac{1}{|\Omega|} \sum_{(i,j,k)\in\Omega} \Big(t_{ijk} - \overbrace{\langle u_i, v_j, w_k \rangle}\Big)^2 + \lambda \|U\|_F^2 + \dots$$

- The problem was partially popularized by the Netflix prize collaborative filtering problem

Matrix Completion

Given a partially obs. matrix



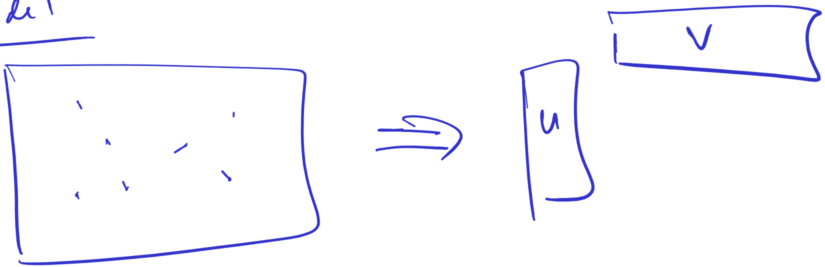
Ω - set of pairs of indices

$a_{ij} \quad \forall\ (i,j) \in \Omega$

recommendation systems

Netflix prize

Model



Objective

$$f(U,V) = \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}\left(a_{ij} - \sum_{r=1}^{R} u_{ir}v_{jr}\right)^2 + \lambda\|U\|_F^2 + \lambda\|V\|_F^2$$

no closed form solution

but $f(U,V)$ is convex if we fix U or V

CP Tensor Completion Gradient and Hessian

$\frac{\partial}{\partial u_i}\langle u_i, v_j, w_k\rangle = v_j * w_k$

- The gradient of the tensor completion objective function is sparsified according to the set of observed entries

$$f(u,v,w) = \frac{1}{|\Omega|}\sum_{(i,j,k)\in\Omega}\left(t_{ijk} - \langle u_i, v_j, w_k\rangle\right)^2 + \lambda\|U\|_F^2 + \cdots$$

$$\frac{\partial f}{\partial u_i} = \frac{1}{|\Omega|}\sum_{(i,j,k)\in\Omega_i}\left(\left(t_{ijk} - \langle u_i, v_j, w_k\rangle\right)(v_j * w_k)\right) + 2\lambda u_i$$

$$\Omega_i = \{(i,j,k) : (i,j,k)\in\Omega\}$$

- ALS for tensor decomposition solves quadratic optimization problem for each row of each factor matrix, in the completion case, Newton's method on these subproblems yields different Hessians

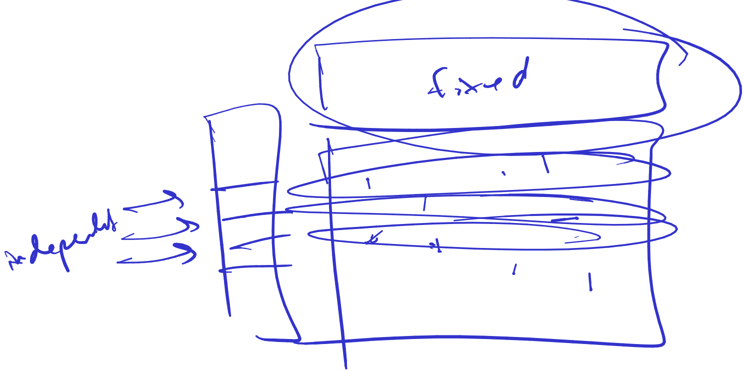
$$H_F^{(u_i)}(u,v,w) = \frac{1}{\Omega}\sum_{(i,j,k)\in\Omega_i}\underbrace{(v_j * w_k)(v_j * w_k)^T}_{v_j v_j^T + w_k w_k^T} + 2\lambda I$$

$$\Omega_i = \text{all } (j,k) \in [1,n]^2$$

$$U = \frac{1}{\Omega} \sum_j \sum_k \underline{v_j v_j^T} * w_k w_k^T$$

$$U = \frac{1}{\Omega} \left(\sum_j v_j v_j^T \right) \left(\sum_k w_k w_k^T \right)$$

$$= \frac{1}{\Omega} \underline{V^T V} * \underline{W^T W} = (V \odot W)^T (V \odot W)$$


fixed
independent

$$\sum_{(j,k)\in\Omega_i} t_{ijk} v_{jr} w_{kr}$$

define $\bar{T}$ so that $\bar{T}_{ijk} = t_{ijk}$ $\forall\ (i,j,k) \in \Omega$

$= 0$ otherwise

then

$$\underbrace{\sum_j \sum_k \bar{T}_{ijk} v_{jr} w_{kr}}_{\text{sparse MTTKRP}} =$$

Methods for CP Tensor Completion

$\mathcal{T} \in \mathbb{R}^{n \times n \times n}$ $|\Omega| = nnz$

- ALS for tensor completion with CP decomposition incurs additional cost

1. To optimize U need to form n Hessians, and solve with $H^{(u_i)} \approx \sum_{(j,k)\in\Omega_i} v_j v_j^T * w_k w_k^T + 2\lambda I$

2. for RHS need MTTKRP with $\mathcal{T}$ (observed entries of $\mathcal{T}$)
(1) has cost $O(\sum_i |\Omega_i| R^2) = O(nnz R^2)$ + solves $O(nR^3)$ (2) $O(nnz \cdot R)$

- Alternative methods for tensor completion include coordinate descent and stochastic gradient descent

  - Coord. desc. minimize for some u_{ir} instead of u_i at a time (less progress / subproblem), but updates cheaper
  - SGD: consider subgradients

consider (random) (i,j,k) $\rightarrow$ $u_i^{(new)} = u_i - \frac{1}{|\Omega|} \cdot \eta \cdot (t_{ijk} - \langle u_i, v_j, w_k \rangle)(v_j * w_k)$

Coordinate Descent for CP Tensor Completion

- Coordinate descent avoids the need to solve linear systems of equations

$$f(u, v, w) = \frac{1}{|\Omega|} \sum_{(i,j,k) \in \Omega} \left(t_{ijk} - \langle u_i, v_j, w_k \rangle \right)^2 + \lambda \|U\|_F^2 + \dots$$

$$\langle u_i, v_j, w_k \rangle = \sum_{r=1}^{R} u_{ir} v_{jr} w_{kr}$$

$$\frac{\partial f}{\partial u_{ir}} = \frac{1}{|\Omega|} \sum_{(i,j,k) \in \Omega_i} \left(t_{ijk} - \langle u_i, v_j, w_k \rangle \right) \left(v_{jr} w_{kr} \right) + 2\lambda u_{ir}$$

$$\rho_{ijk}^{(r)} = t_{ijk} - \langle u_i, v_j, w_k \rangle + u_{ir} v_{jr} w_{kr} = t_{ijk} - \sum_{r'=1, r' \neq r}^{R} u_{ir'} v_{jr'} w_{kr'}$$

$$= \frac{1}{|\Omega|} \sum_{(i,j,k) \in \Omega_i} \left(\rho_{ijk}^{(r)} + u_{ir} v_{jr} w_{kr} \right) \left(v_{jr} w_{kr} \right) + 2\lambda u_{ir}$$

$$u_{ir} = \frac{1}{|\Omega|2\lambda} \sum_{(j,k)\in\Omega_i} \rho^{(r)}_{ijk} v_{jr} w_{kr} + u_{ir} \frac{1}{|\Omega|2\lambda} \sum_{(j,k)\in\Omega_i} v_{jr} w_{kr}$$

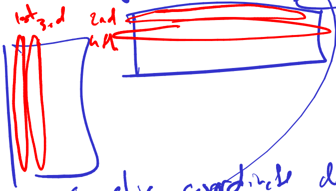
residual-like term

$$= \ldots \frac{\sum_{(j,k)\in\Omega_i} \rho^{(r)}_{ijk} v_{jr} w_{kr}}{1 + \frac{1}{|\Omega|2\lambda} \sum_{(j,k)\in\Omega_i} v_{jr} w_{kr}}$$

CCD

CCD++

diff. ordering choices



or

"ALS" ordering

cyclic coordinate descent

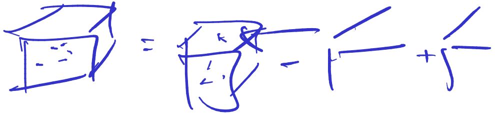
Suppose we've updated u_{ir} $\forall i$ (one column)

with $\forall (i,j,k) \in \Omega$ $\Big|$ $\underline{\rho^{(r)}_{ijk}} = \left(t_{ijk} - \langle u_{ir}, v_{jr}, w_{kr}\rangle + u_{ir} v_{jr} w_{kr}\right)$

next we may consider $\rho^{(r+1)}_{ijk} = \rho^{(r)}_{ijk} - u_{ir} v_{jr} w_{kr} + u_{i,r+1} v_{j,r+1} w_{k,r+1}$

CP cost per col-update is $O(|\Omega|)$ so for all factors/cols!

$O(|\Omega| R)$



$O(|\Omega|)$ cost

Sparse Tensor Contractions

- Tensor completion and sparse tensor decomposition require operations on sparse tensors

- Sparse tensor contractions often correspond to products of *hypersparse* matrices, i.e., matrices with mostly zero rows

Sparse Tensor Formats

- The overhead of transposition, and non-standard nature of the arising sparse matrix products, motivates sparse data structures for tensors that are suitable for tensor contractions of interest

- The *compressed sparse fiber (CSF)* format provides an effective representation for sparse tensors

Operations in Compressed Format

- CSF permits efficient execution of important sparse tensor kernels
  - Analogous to CSR format, which enables efficient implementation of the sparse matrix vector product
  - where `row[i]` stores a list of column indices and nonzeros in the ith row of $\boldsymbol{A}$

```
for i in range(n):
  for (a_ij,j) in row[i]:
    y[i] += a_ij * x[j]
```

  - In CSF format, a multilinear function evaluation $\boldsymbol{f}^{(\boldsymbol{\mathcal{T}})}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{T}_{(1)}(\boldsymbol{x} \odot \boldsymbol{y})$ can be implemented as

```
for (i,T_i) in T_CSF:
  for (j,T_ij) in T_i:
    for (k,t_ijk) in T_ij:
      z[i] += t_ijk * x[j] * y[k]
```

MTTKRP in Compressed Format

- MTTKRP and CSF pose additional implementation opportunities and challenges
  - MTTKRP $u_{ir} = \sum_{j,k} t_{ijk} v_{jr} w_{kr}$ can be implemented by adding a loop over r to our code for $\boldsymbol{f}^{(\mathcal{T})}$, but would then require $3mr$ operations if m is the number of nonzeros in $\mathcal{T}$, can reduce to $2mr$ by amortization

```
for (i,T_i) in T_CSF:
  for (j,T_ij) in T_i:
    for r in range(R):
      f_ij = 0
      for (k,t_ijk) in T_ij:
        f_ij += t_ijk * w[k,r]
      u[i,r] = f_ij * v[j,r]
```

  - However, this amortization is harder (requires storage or iteration overheads) if the index i is a leaf node in the CSF tree
  - Similar challenges in achieving good reuse and obtaining good arithmetic intensity arise in implementation of other kernels, such as TTMc

All-at-once Contraction

- When working with sparse tensors, it is often more efficient to contract multiple operands in an all-at-once fashion

Constrained Tensor Decomposition

- Many applications of tensor decomposition in data science, feature additional structure, which can be enforced by constraints

$$\min_{U,V,W \in S} \| T - [\![U,V,W]\!] \|_F$$

$$\Downarrow$$

$U,V,W \geq 0$ nonnegativity

$U^T U \approx I$ (orthogonality or near orthogonality)

$\| u_i - u_{i+1} \| \leq C$ (continuity)

time $\downarrow$

Nonnegative Tensor Factorization

- *Nonnegative tensor factorization (NTF)*, such as CP decomposition with $\mathcal{T} \geqslant 0$ and $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W} \geqslant 0$ are widespread and a few classes of algorithms have been developed

proximal optimization method

$$\min_{u \in S} \left(f(u) + \tfrac{1}{2}\|u - x\|_2^2 \right)$$

$\mathbb{R}_+^n$

optimal (outside S)

Nonnegative Matrix Factorization

- NTF algorithms with alternating updates have a close correspondence with alternating update algorithms for *Nonnegative matrix factorization (NMF)*

Coordinate Descent for NMF and NTF

- Coordinate descent gives optimal closed-form updates for variables in NMF and NTF

$$\min_{u_i, v \in \mathbb{R}_+^{n \times R}} f(u,v) = \| T - uv^T \|_2^2$$

$$\frac{\partial f}{\partial u_i} = (T - uv^T) v_i = 0 = P v_i - u_i v_i^T v_i \Rightarrow$$

$$P = T - uv^T + u_i v_i^T \qquad u_i = \frac{P v_i}{v_i^T v_i}$$

$$u_i = \frac{P v_i}{v_i^T v_i} \qquad u_i = \left| \frac{P v_i}{v_i^T v_i} \right|_+ \qquad (|x|_+)_{(i)} = \begin{cases} x_i & \text{if } x_i \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Generalized Tensor Decomposition

- Aside from addition of constraints, the objective function may be modified by using different elementwise loss functions
- Some loss function admit ALS-like algorithms, while others may require gradient-based optimization