Fixed-Sparsity Matrix Approximation from Matrix-Vector Products

Noah Amsel, Tyler Chen, Feyza Duman Keles,
Diana Halikias, Cameron Musco, Christopher Musco





1. Problem

Structured Matrix Approximation

Find the best approximation from some structured class:

$$\min_{\hat{\mathbf{A}} \in \mathcal{S}} \|\mathbf{A} - \hat{\mathbf{A}}\|$$

- $\mathcal{S}$ is rank-k matrices → truncated SVD

Fixed-Pattern Sparse Approximation

Let $\mathbf{S} \in \{0,1\}^{n \times d}$ be a sparsity pattern:

$$\underset{\hat{\mathbf{A}} = \mathbf{S} \circ \hat{\mathbf{A}}}{\operatorname{argmin}} \|\mathbf{A} - \hat{\mathbf{A}}\|_{\mathsf{F}} = \mathbf{A} \circ \mathbf{S}$$

$$\mathbf{S} = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}$$ → just extract the diagonal

- Banded, block diagonal, etc.

Matvec Access Model

- Queries: $\mathbf{x}_1, \ldots, \mathbf{x}_m \mapsto \mathbf{A}\mathbf{x}_1, \ldots, \mathbf{A}\mathbf{x}_m$
- E.g. $\mathbf{A} = \mathbf{B}^{-1}$
- (Adaptive? Transpose queries? … You'll see)

Approximate Structured Matrix Approximation

- Compete with the best structured matrix approximation
- Find $\tilde{\mathbf{A}} \in \mathcal{S}$ such that

$$\|\mathbf{A} - \tilde{\mathbf{A}}\| \leq (1+\epsilon) \min_{\hat{\mathbf{A}} \in \mathcal{S}} \|\mathbf{A} - \hat{\mathbf{A}}\|$$

- SVD → RandSVD

Our Problem

"Approximate sparse approximation in the matvec access model"

Given

- $\mathbf{S} \in \{0,1\}^{n \times d}$
- matvec access to $\mathbf{A} \in \mathbb{R}^{n \times d}$

find sparse $\tilde{\mathbf{A}} = \mathbf{S} \circ \tilde{\mathbf{A}}$ such that

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_\mathsf{F} \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{S} \circ \mathbf{A}\|_\mathsf{F}$$

What this is *not*

- Exact recovery
  - An exactly diagonal matrix can be recovered exactly with one matvec
  - Easier
- Compressed sensing (matrix version)
  - Unknown support
  - Harder

2. Upper Bound

Idea

- Sketch $\mathbf{A}$ with m Gaussians

$$\begin{bmatrix} & & \\ & \mathbf{A} & \\ & & \end{bmatrix} \begin{bmatrix} | & & | \\ \mathbf{g}_1 & \cdots & \mathbf{g}_m \\ | & & | \end{bmatrix} = \begin{bmatrix} & \mathbf{Z} & \end{bmatrix}$$

- Solve a least squares problem for each row

$$\begin{bmatrix} a_{11} & ? & ? & ? \end{bmatrix} \begin{bmatrix} g_{11} & & g_{1m} \\ \vdots & \cdots & \vdots \\ g_{d1} & & g_{dm} \end{bmatrix} = \begin{bmatrix} z_{11} & \cdots & z_{1m} \end{bmatrix}$$

Idea

- Sketch $\mathbf{A}$ with m Gaussians

$$\begin{bmatrix} & & \\ & \mathbf{A} & \\ & & \end{bmatrix} \begin{bmatrix} | & & | \\ \mathbf{g}_1 & \cdots & \mathbf{g}_m \\ | & & | \end{bmatrix} = \begin{bmatrix} & \mathbf{Z} & \end{bmatrix}$$

- Solve a least squares problem for each row

$$a_{11} \, [g_{11} \quad \cdots \quad g_{1m}] + [? \quad ? \quad ?] \, \mathbf{G}' = [z_{11} \quad \cdots \quad z_{1m}]$$

Idea

- Sketch $\mathbf{A}$ with m Gaussians

$$\begin{bmatrix} & & \\ & \mathbf{A} & \\ & & \end{bmatrix} \begin{bmatrix} | & & | \\ \mathbf{g}_1 & \cdots & \mathbf{g}_m \\ | & & | \end{bmatrix} = \begin{bmatrix} & \mathbf{Z} & \end{bmatrix}$$

- Solve a least squares problem for each row

$$[a_{11} \quad a_{21}] \begin{bmatrix} g_{11} & \cdots & g_{1m} \\ g_{21} & \cdots & g_{2m} \end{bmatrix} + [? \quad ?]\, \mathbf{G}' = [z_{11} \quad \cdots \quad z_{1m}]$$

Upper bound

If $\mathbf{S}$ has $\leq s$ non-zeros per row, then we need only $m = O\left(\frac{s}{\epsilon}\right)$ matvecs to solve w.h.p.

- Dimension free!
- Non-adaptive queries!
- Generalizes Hutchinson's diagonal estimator
  - [Batson & Nakatsukasa '22] [Dharangutte and Musco '23]
- Coloring / probing methods [Curtis Powell Reid '74] [Frommer Schimmel Schweitzer '21] [Schäfer Owhadi '21]
  - Worse even for exact case with doubly sparse $\mathcal{S}$: $m = \Omega(s^2)$
  - Beats us by $(m-s)/m$ for some banded matrices

Fact: if $\mathbf{G} \in \mathbb{R}^{m \times s}$ and $m \geq s+2$ then $\mathbb{E}\left[\|\mathbf{G}^\dagger\|_{\mathsf{F}}^2\right] = \dfrac{s}{m-s-1}$ cf. [HMT 11]

3. Lower Bound

Hard Instance

Let

- $\mathbf{G} \in \mathbb{R}^{d \times d}$ have iid Gaussian entries
- $\mathbf{A} = \mathbf{G}^\top \mathbf{G}$ (Wishart)
  - Linear Regression, PCA, trace estimation
  - [Braverman et al. '20] [Simchowitz, Alaoui, Recht ’18] [Jiang et al. ’21]
- $\mathbf{S}$ has between s/2 and s entries per row and column (e.g., block diagonal, banded)

Properties

- Symmetric, psd
- $\mathbf{I}$ is special case
- Turns out, adaptive queries can’t help much

A Wishart given matvec queries is still Wishart

Query $\mathbf{G}^\top \mathbf{G} \in \mathbb{R}^{d \times d}$ with m adaptive matvec queries

Then there exists $\mathbf{\Delta} \in \mathbb{R}^{d \times d}$ and orthonormal $\mathbf{V}$ s.t. the posterior distribution is

$$\mathbf{G}^\top \mathbf{G} \sim \mathbf{V} \left(\mathbf{\Delta} + \begin{bmatrix} \overset{m}{\cdot} & \overset{(d-m)}{\cdot} \\ \cdot & \mathbf{G}_2^\top \mathbf{G}_2 \end{bmatrix} \right) \mathbf{V}^\top$$

[Braverman, Hazan, Simchowitz, Woodworth ’20], used in several others

Anti-concentration of Wishart entries

- (From Berry-Esseen and anti-concentration of Gaussians)
- Let $\mathbf{G} \in \mathbb{R}^{k \times k}$ have Gaussian entries
- Impossible to accurately estimate $\mathbf{e}_i^\top \mathbf{G}^\top \mathbf{G} \mathbf{e}_j$ to accuracy better than $\sqrt{k}$

Anti-concentration of (rotated) Wishart entries

- (From Berry-Esseen and anti-concentration of Gaussians)
- Let $\mathbf{G} \in \mathbb{R}^{k \times k}$ have Gaussian entries
- Impossible to accurately estimate $\mathbf{u}^\top \mathbf{G}^\top \mathbf{G} \mathbf{v}$ to accuracy better than $\sqrt{k}$

Lower Bound

Let

- $\mathbf{G} \in \mathbb{R}^{d \times d}$ have iid Gaussian entries
- Let $\mathbf{A} = \mathbf{G}^\top \mathbf{G}$
- Let $\mathbf{S}$ have $\Theta(s)$ entries per row/column (e.g., block diagonal)

Then:

$m = \Omega\left(\frac{s}{\epsilon}\right)$ queries are needed to achieve $(1+\epsilon)$ error w.p. $\geq 5\,\%$

even if the queries are adaptive

In conclusion

The matvec query complexity of approximate sparse approximation is $\Theta\left(s/\epsilon\right)$

Open questions

- Beyond Frobenius norm
- Combining with “coloring methods”
- Other important classes: sparse + low rank, *hierarchical*, …

Applications

- $f(\mathbf{A})$ where $\mathbf{A}$ is banded [Park and Nakatsukasa 2023]
- $[\mathrm{Cov}(\mathbf{X})]^{-1}$ where $\mathbf{X}$ is drawn from a Gaussian Markov random field

Runtime

- Naively, must solve n least squares problems of size $m \times s$ so $O(nms^2)$
- For many sparsity patterns, you can reuse most work from the ith system to solve the $(i+1)$th system fast
- Embarrassingly parallel

Pros/cons of Coloring Methods

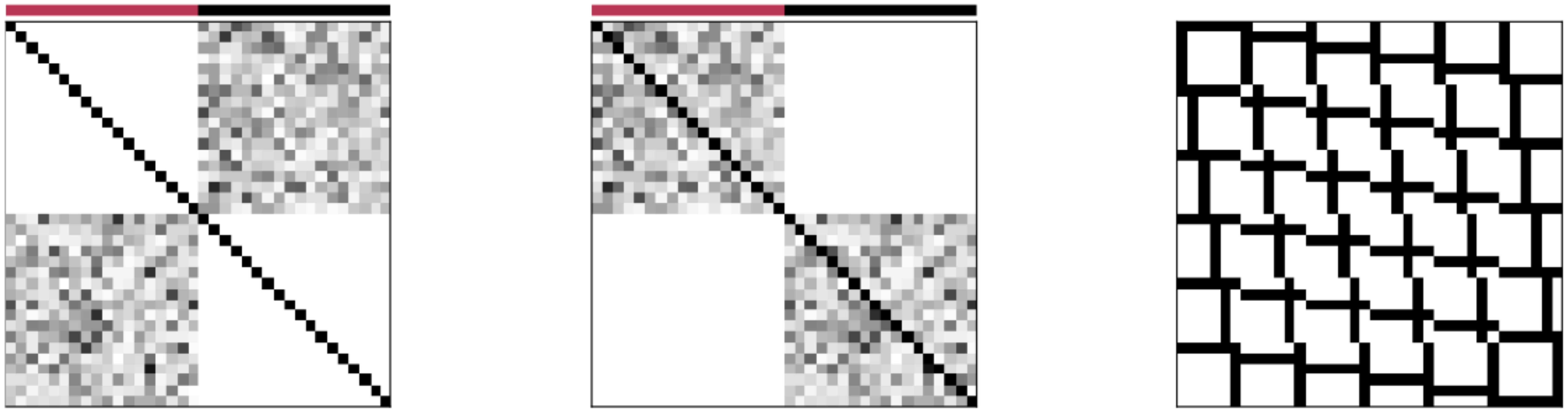

Figure 1: Left*: Visualization of a matrix describedin* Section 4.2 *for which* Algorithm 1 *is not the best method for recovering the diagonal (intensity indicates magnitude of entries of* **A***). In particular, the diagonal of the matrix can be recovered using exactly 2 queries, while* Algorithm 1 *will require many queries to overcome the large noise in the off-diagonal blocks.* Middle*: Visualization of a matrix for which using the same colorings as the matrix on the left panel will not help.* Right*: Visualization of the hard sparsity pattern described in* Section 4.3 *with* $k = 10$*. Here black pixels correspond to one and white pixels to zero. Note that while each row and column of the matrix has only* $O(k)$ *nonzeros, each pair of the* k^2 *columns has overlapping support.*