A Matrix Chernoff Bound for Markov Chains and its Application to Co-occurrence Matrices

Jiezhong Qiu, Chi Wang, Ben Liao, Richard Peng, Jie Tang
The Application to Co-occurrence Matrices

<table>
<thead>
<tr>
<th>counts</th>
<th>i</th>
<th>like</th>
<th>enjoy</th>
<th>deep</th>
<th>learning</th>
<th>NLP</th>
<th>flying</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>like</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>enjoy</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>deep</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>learning</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NLP</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>flying</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

NLP
(LDA, Word2vec, Glove)

Graph Learning
(DeepWalk, node2vec, metapath2vec)

Recommendation System
(Pin2Vec, Item2vec)

Reinforcement Learning
(Act2Vec)

Hidden Markov Models
(Emission Co-occurrence)
Theorem (Chernoff Bound, 1952): If $X_1, X_2, \ldots, X_k$ are independent zero-mean scalar-valued random variables with $|X_i| \leq 1$. Then for $\epsilon \in (0, 1)$

$$\Pr\left(\left|\frac{1}{k} \sum_{i=1}^{k} X_i \right| \geq \epsilon\right) \leq 2\exp(-k\epsilon^2/4)$$

The sample mean $\frac{1}{k} \sum_{i=1}^{k} X_i$ falls into this area with high prob.
A Matrix Chernoff Bound for Markov Chains

Independence
Markov Dependence

Scalar-valued
Random Variables
Matrix-valued
Random Variables

Sample Mean Matrix

\[ \frac{1}{2} (f(X_1) + f(X_2)) \]
A Matrix Chernoff Bound for Markov Chains

Independence
Markov Dependence

Scalar-valued
Random Variables
Matrix-valued
Random Variables

Sample Mean Matrix

\[
\frac{1}{3} \left( f(X_1) + f(X_2) + f(X_3) \right)
\]
A Matrix Chernoff Bound for Markov Chains

Independence
Markov Dependence

Scalar-valued
Random Variables
Matrix-valued
Random Variables

Sample Mean Matrix

\[ \frac{1}{4} (f(X_1) + f(X_2) + f(X_3) + f(X_4)) \]

Eigenvalues of the sample mean matrix

\[ \frac{1}{k} \sum_{i=1}^{k} f(X_i) \]

falls into this area with high prob.

Left Tail
\[ \mathbb{P}[\leq -\epsilon] \]

Right Tail
\[ \mathbb{P}[\geq \epsilon] \]
A Matrix Chernoff Bound for Markov Chains

\[ \mathbb{P} \left[ \lambda_{\min} \left( \frac{1}{k} \sum_{i=1}^{k} f(X_i) \right) \leq -\epsilon \right] \quad \text{and} \quad \mathbb{P} \left[ \lambda_{\max} \left( \frac{1}{k} \sum_{i=1}^{k} f(X_i) \right) \geq \epsilon \right] \]

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Chernoff `52</th>
<th>Tropp`12</th>
<th>GLSS`18</th>
<th>Our Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X)</td>
<td>i.i.d scalars</td>
<td>i.i.d matrices</td>
<td>Stationary random walk on an undirected regular graph with spectral expansion (\lambda)</td>
<td>Non-stationary random walk on a regular Markov chain with spectral expansion (\lambda)</td>
</tr>
<tr>
<td>(f(X))</td>
<td>(X)</td>
<td>(X)</td>
<td>(d \times d) matrix</td>
<td>(d \times d) matrix</td>
</tr>
<tr>
<td>tail prob.</td>
<td>(\exp(-\Omega(k\epsilon^{-2})))</td>
<td>(\exp(-\Omega(k\epsilon^{-2})))</td>
<td>(\exp(-\Omega(k(1 - \lambda)\epsilon^{-2})))</td>
<td>(\exp(-\Omega(k(1 - \lambda)\epsilon^{-2})))</td>
</tr>
</tbody>
</table>
A Matrix Chernoff Bound for Markov Chains

Theorem: Let $P$ be an regular Markov chain with state space $[N]$, stationary distribution $\pi$ and spectral expansion $\lambda$. Let $f: [N] \to \mathbb{C}^{d \times d}$ be a matrix-valued function such that

1. $\forall X \in [N], f(X)$ is Hermitian and $\|f(X)\|_2 \leq 1$;
2. $\sum_{X \in [N]} \pi_X f(X) = 0$.

Let $(X_1, X_2, \ldots, X_k)$ denote a $k$-step random walk on $P$ starting from an initial distribution $\phi$. Then for $\epsilon \in (0, 1)$:

$$\mathbb{P} \left[ \lambda_{\min} \left( \frac{1}{k} \sum_{i=1}^{k} f(X_i) \right) \leq -\epsilon \right] \leq \|\phi\|_\pi d^2 \exp(-k(1 - \lambda)\epsilon^2 / 72)$$

$$\mathbb{P} \left[ \lambda_{\max} \left( \frac{1}{k} \sum_{i=1}^{k} f(X_i) \right) \geq \epsilon \right] \leq \|\phi\|_\pi d^2 \exp(-k(1 - \lambda)\epsilon^2 / 72)$$
The Application to Co-occurrence Matrices

<table>
<thead>
<tr>
<th>counts</th>
<th>like</th>
<th>enjoy</th>
<th>deep</th>
<th>learning</th>
<th>NLP</th>
<th>flying</th>
</tr>
</thead>
<tbody>
<tr>
<td>like</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>enjoy</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>deep</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>learning</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NLP</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>flying</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

NLP (LDA, Word2vec, Glove)

Graph Representation Learning (DeepWalk, node2vec, metapath2vec)

Recommendation System (Pin2Vec, Item2vec)

Reinforcement Learning (Act2Vec)

Hidden Markov Models (Emission Co-occurrence)
Co-occurrence Matrix of Sequential Data

Sliding Window 1
\[ X_1 = (1,2,3) \]

\[
C = \frac{1}{4} \begin{pmatrix}
0 & 1 & 1 \\
1 & 0 & 0 \\
1 & 0 & 0
\end{pmatrix}
\]
Co-occurrence Matrix of Sequential Data

Sliding Window 2
\[ X_2 = (2,3,2) \]

\[
C = \frac{1}{2} \left[ \frac{1}{4} \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} + \frac{1}{4} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \end{pmatrix} \right]
\]
Co-occurrence Matrix of Sequential Data

Sliding Window 3

\[ x_3 = (3,2,3) \]

\[ C = \frac{1}{3} \left[ \frac{1}{4} \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} + \frac{1}{4} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \end{pmatrix} + \frac{1}{4} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right] \]
Observation 1:
Let $X_1, X_2, \ldots, X_{L-T}$ be the sequence of sliding windows, and $f$ maps a sliding window to the co-occurrence matrix within this window. The co-occurrence matrix $C$ can be written as the sample mean of $f(X_1), f(X_2), \ldots, f(X_{L-T})$:

$$C = \frac{1}{L-T} \sum_{k=1}^{L-T} f(X_k)$$

Observation 2: If the input sequence $v_1, v_2, \ldots$ is a Markov Chain, then $X_1, X_2, \ldots$ is a Markov Chain, too.
Convergence Rate of Co-occurrence Matrices

- The co-occurrence matrix:
  \[ C = \frac{1}{L-T} \sum_{k=1}^{L-T} f(X_k) \]
- The asymptotic expectation of \( C \) (denote \( \Pi = \text{diag}(\pi) \)):
  \[
  \mathbb{A}E[C] = \lim_{L \to +\infty} \mathbb{E}[C] = \sum_{r=1}^{T} \frac{1}{2T} (\Pi P^r + (\Pi P^r)^T)
  \]

Theorem: Let \( P \) be a regular Markov chain with state space \([n]\), stationary distribution \( \pi \) and mixing time \( \tau \). Let \((v_1, \cdots, v_L)\) be a \( L \)-step random walk on \( P \) starting from a distribution \( \phi \). Given \( \epsilon \in (0, 1) \), the probability that the co-occurrence matrix \( C \) deviates from its asymptotic expectation \( \mathbb{A}E[C] \) (in 2-norm) is bounded by:

\[
P(\|C - \mathbb{A}E[C]\|_2 \geq \epsilon) \leq 2(\tau + T)\|\phi\|_\pi n^2 \exp \left( -\frac{\epsilon^2 (L - T)}{576(\tau + T)} \right)
\]

Roughly, one needs \( L = O(\tau(\log n + \log \tau)/\epsilon^2) \) samples to guarantee good estimation to the co-occurrence matrix.
## Comparison

<table>
<thead>
<tr>
<th></th>
<th>$X$</th>
<th>$f(X)$</th>
<th>Tail Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chernoff `52</td>
<td>i.i.d scalars</td>
<td>$X$</td>
<td>$\exp(-\Omega(ke^{-2}))$</td>
</tr>
<tr>
<td>Tropp`12</td>
<td>i.i.d matrices</td>
<td>$X$</td>
<td>$d\exp(-\Omega(ke^{-2}))$</td>
</tr>
<tr>
<td>GLSS`18</td>
<td>Stationary random walk on an undirected regular graph with spectral expansion $\lambda$</td>
<td>$d \times d$ matrix</td>
<td>$d\exp(-\Omega(k(1 - \lambda)e^{-2}))$</td>
</tr>
<tr>
<td>Ours</td>
<td>Non-stationary random walk on a regular Markov chain with spectral expansion $\lambda$</td>
<td>$d \times d$ matrix</td>
<td>$d\exp(-\Omega(k(1 - \lambda)e^{-2}))$</td>
</tr>
<tr>
<td>HKS`15</td>
<td>Size-1 sliding windows on a reversible Markov chain on $[n]$ with mixing time $\tau$</td>
<td>Co-occurrence matrix within window</td>
<td>$\tau n \exp(-\Omega(L e^{-2}/\tau))$</td>
</tr>
<tr>
<td>Ours</td>
<td>Size-$T$ sliding windows on a regular Markov chain on $[n]$ with mixing time $\tau$</td>
<td>Co-occurrence matrix within window</td>
<td>$(\tau + T) n \exp(-\Omega(L e^{-2}/(\tau + T)))$</td>
</tr>
</tbody>
</table>
Numerical Experiments

Figure 1: The convergence rate of co-occurrence matrices on Barbell graph, winning streak chain, BlogCatalog graph, and random graph (in log-log scale). The $x$-axis is the trajectory length $L$ and the $y$-axis is the approximation error $\|C - \mathbb{E}[C]\|_2$. Each experiment contains 64 trials, and the error bar is presented.
Thanks!