Adversarially-Trained Nonnegative Matrix Factorization

Ting Cai     Vincent Y. F. Tan     Cédric Févotte

ICASSP (Singapore), May 2022
Given a nonnegative data matrix $V \in \mathbb{R}_{+}^{F \times N}$, approximate $V$ as

$$V \approx WH$$

where $W \in \mathbb{R}_{+}^{F \times K}$ (basis) and $H \in \mathbb{R}_{+}^{K \times N}$ (coefficient) matrices.
Given a **nonnegative** data matrix \( V \in \mathbb{R}^{F \times N}_+ \), approximate \( V \) as

\[
V \approx WH
\]

where \( W \in \mathbb{R}^{F \times K}_+ \) (basis) and \( H \in \mathbb{R}^{K \times N}_+ \) (coefficient) matrices.
Given a nonnegative data matrix $V \in \mathbb{R}^{F \times N}_+$, approximate $V$ as

$$V \approx WH$$

where $W \in \mathbb{R}^{F \times K}_+$ (basis) and $H \in \mathbb{R}^{K \times N}_+$ (coefficient) matrices.

One typically solves

$$\min_{W, H \geq 0} D(V | WH) \quad \text{e.g.} \quad D(V | WH) = \|V - WH\|_F^2,$$

where $A \geq 0$ means that all the entries of $A$ are nonnegative.
Motivation and Main Contribution

- **Adversarial training** (Goodfellow et al., 2015; Madry et al., 2018; Tramèr et al., 2018)

\[
\min_{\theta \in \Theta} \quad \max_{x' : \|x - x'\| \leq \epsilon} \quad \frac{1}{n} \sum_{i=1}^{n} \text{Loss}(f_{\theta}(x'_i), y_i).
\]

Improve the predictive performance of NMF using adversarial training for matrix completion tasks. Derive efficient algorithms for updating the adversary and \((W, H)\). Demonstrate the superior predictive performance of adversarially-trained NMF or AT-NMF over other methods on matrix completion tasks for three benchmark datasets.
Motivation and Main Contribution

- **Adversarial training** (Goodfellow et al., 2015; Madry et al., 2018; Tramèr et al., 2018)

\[
\min_{\theta \in \Theta} \max_{x':\|x-x'\| \leq \epsilon} \frac{1}{n} \sum_{i=1}^{n} \text{Loss}(f_{\theta}(x'_i), y_i).
\]

- Improve the predictive performance of NMF using adversarial training for matrix completion tasks.

- Derive efficient algorithms for updating the adversary and \((W,H)\).

- Demonstrate the superior predictive performance of adversarially-trained NMF or AT-NMF over other methods on matrix completion tasks for three benchmark datasets.
Motivation and Main Contribution

- **Adversarial training** (Goodfellow et al., 2015; Madry et al., 2018; Tramèr et al., 2018)

  \[ \min_{\theta \in \Theta} \max_{\|x - x'\| \leq \epsilon} \left( \frac{1}{n} \sum_{i=1}^{n} \text{Loss} \left( f_{\theta}(x'_i), y_i \right) \right) \]

  perturbation of features

- Improve the predictive performance of NMF using adversarial training for matrix completion tasks.

- Derive **efficient algorithms** for updating the adversary and \((W, H)\).
Motivation and Main Contribution

- **Adversarial training** (Goodfellow et al., 2015; Madry et al., 2018; Tramèr et al., 2018)

\[
\begin{align*}
\min_{\theta \in \Theta} & \quad \max_{x': \|x-x'\| \leq \epsilon} \quad \frac{1}{n} \sum_{i=1}^{n} \text{Loss}(f_{\theta}(x'_i), y_i).
\end{align*}
\]

perturbation of features

- Improve the predictive performance of NMF using adversarial training for matrix completion tasks.

- Derive **efficient algorithms** for updating the adversary and \((W, H)\).

- Demonstrate the **superior predictive performance** of adversarially-trained NMF or AT-NMF over other methods on matrix completion tasks for three benchmark datasets.
Formulation of AT-NMF

Consider an adversary that adds an arbitrary matrix $\mathbf{R} \in \mathbb{R}^{F \times N}_+$ to $\mathbf{V}$ to maximize the divergence between $\mathbf{V}$ and $\mathbf{W}\mathbf{H}$, AT-NMF is formulated as

$$\min_{\mathbf{W}, \mathbf{H} \geq 0} \max_{\mathbf{R} \in \mathcal{R}} \| \mathbf{V} + \mathbf{R} - \mathbf{W}\mathbf{H} \|_F^2$$

where the constraint set

$$\mathcal{R} = \{ \mathbf{R} : \| \mathbf{R} \|_F^2 \leq \epsilon, \mathbf{V} + \mathbf{R} \geq \mathbf{0} \}$$
Formulation of AT-NMF

Consider an adversary that adds an arbitrary matrix $R \in \mathbb{R}^{F \times N}_+$ to $V$ to maximize the divergence between $V$ and $WH$, AT-NMF is formulated as

$$\min_{W,H \geq 0} \max_{R \in \mathcal{R}} \|V + R - WH\|_F^2$$

where the constraint set

$$\mathcal{R} = \{ R : \|R\|_F^2 \leq \epsilon, V + R \geq 0 \}$$

$\epsilon > 0$ is a constant indicating the adversary's power.
Consider an adversary that adds an arbitrary matrix $R \in \mathbb{R}^{F \times N}_+$ to $V$ to maximize the divergence between $V$ and $WH$, AT-NMF is formulated as

$$\min_{W,H \geq 0} \max_{R \in \mathcal{R}} \|V + R - WH\|_F^2$$

where the constraint set

$$\mathcal{R} = \{R : \|R\|_F^2 \leq \epsilon, V + R \geq 0\}$$

$\epsilon > 0$ is a constant indicating the adversary's power.

To relax the problem, dualize the constraint $\|R\|_F^2 \leq \epsilon$ with Lagrange multiplier $\lambda > 0$, AT-NMF becomes

$$\min_{W,H \geq 0} \max_{R:V+R \geq 0} \|V + R - WH\|_F^2 - \lambda \|R\|_F^2.$$
AT-NMF Algorithm (Update of $\mathbf{R}$)

Let $\hat{\mathbf{V}} = \mathbf{W} \mathbf{H}$, the inner maximization problem can be rewritten as a minimization problem as

$$R^* = \arg \min_{\mathbf{R}} : \quad \mathbf{V} + \mathbf{R} \geq 0$$

$$-\| \mathbf{V} + \mathbf{R} - \hat{\mathbf{V}} \|_2^2 + \lambda \| \mathbf{R} \|_2^2$$

Objective separates into $FN$ independent terms $g(\mathbf{R}) = \sum f_n, n \left[ -\left( v_{fn} + r_{fn} - \hat{v}_{fn} \right)^2 + \lambda r_{fn}^2 \right]$. Suffices to minimize each term inside over $r_{fn}$. By re-arranging:

$$\min_{r_{fn}} : \quad v_{fn} + r_{fn} \geq 0$$

$$(\lambda - 1) r_{fn}^2 - 2 r_{fn} (v_{fn} - \hat{v}_{fn})$$

Can be solved in closed-form. For $\lambda \in [0, 1]$, $r_{fn} = \infty$; for $\lambda > 1$, $R^* = \max \{ \mathbf{V} - \hat{\mathbf{V}} \lambda - 1, -\mathbf{V} \}$. 

Let $\hat{V} = WH$, the inner maximization problem can be rewritten as a minimization problem as

$$
\mathbf{R}^* = \arg \min_{\mathbf{R}: \mathbf{V} + \mathbf{R} \geq 0} -\|\mathbf{V} + \mathbf{R} - \hat{V}\|_F^2 + \lambda \|\mathbf{R}\|_F^2
$$
Let $\hat{V} = WH$, the inner maximization problem can be rewritten as a minimization problem as

$$\mathbf{R}^* = \arg \min_{\mathbf{R} : \mathbf{V} + \mathbf{R} \geq 0} -\| \mathbf{V} + \mathbf{R} - \hat{\mathbf{V}} \|^2_F + \lambda \| \mathbf{R} \|^2_F$$

Objective separates into $FN$ independent terms

$$g(\mathbf{R}) = \sum_{f,n} \left[ -\left( v_{fn} + r_{fn} - \hat{v}_{fn} \right)^2 + \lambda r_{fn}^2 \right].$$
AT-NMF Algorithm (Update of $\mathbf{R}$)

- Let $\hat{\mathbf{V}} = \mathbf{WH}$, the inner maximization problem can be rewritten as a minimization problem as

$$
\mathbf{R}^* = \arg\min_{\mathbf{R}: \mathbf{V} + \mathbf{R} \geq 0} -\|\mathbf{V} + \mathbf{R} - \hat{\mathbf{V}}\|_F^2 + \lambda \|\mathbf{R}\|_F^2
$$

- Objective separates into $FN$ independent terms

$$
g(\mathbf{R}) = \sum_{f,n} \left[ -(v_{fn} + r_{fn} - \hat{v}_{fn})^2 + \lambda r_{fn}^2 \right].
$$

- Suffices to minimize each term inside over $r_{fn}$. By re-arranging:

$$
\min_{r_{fn}: v_{fn} + r_{fn} \geq 0} (\lambda - 1) r_{fn}^2 - 2r_{fn}(v_{fn} - \hat{v}_{fn})
$$

- Can be solved in closed-form. For $\lambda \in [0, 1]$, $r_{fn} = \infty$; for $\lambda > 1$,

$$
\mathbf{R}^* = \max \left\{ \frac{\mathbf{V} - \hat{\mathbf{V}}}{\lambda - 1}, -\mathbf{V} \right\}
$$
AT-NMF Algorithm (Update of \((\mathbf{W}, \mathbf{H})\))

After update of \(\mathbf{R}\) to \(\mathbf{R}^*\),

\[
\arg\min_{\mathbf{W}, \mathbf{H}} \|\mathbf{V} + \mathbf{R}^*\|_2 =: \mathbf{U} - \mathbf{WH} - \lambda \|\mathbf{R}\|_2^F
\]

fix \(\mathbf{U} := \mathbf{V} + \mathbf{R}^*\).

Use majorization-minimization (MM) (Hunter and Lange, 2000) to update

\(\mathbf{H} \leftarrow \mathbf{H} \cdot \mathbf{W}^\top \mathbf{U} \mathbf{W}^\top \mathbf{WH}\) and

\(\mathbf{W} \leftarrow \mathbf{W} \cdot \mathbf{U} H^\top \mathbf{WHH}^\top\)

Initialization of \((\mathbf{W}, \mathbf{H})\)

Sample each entry independently from Half-Normal distribution (with variance parameter \(\gamma = 1\));

Run 5 standard MM steps on \(\mathbf{V}\) to obtain \(\mathbf{W}_{\text{init}}\) and \(\mathbf{H}_{\text{init}}\).

Termination (described in paper)
AT-NMF Algorithm (Update of \((W, H)\))

- After update of \(R\) to \(R^*\),

\[
\arg \min_{W, H \geq 0} \left\| V + R^* - WH \right\|_F^2 - \lambda \left\| R \right\|_F^2 =: U
\]

\(\text{fix } U := V + R^*\).
AT-NMF Algorithm (Update of \((W, H)\))

- After update of \(R\) to \(R^*\),

\[
\arg\min_{W,H \geq 0} \| V + R^* - WH \|_F^2 - \lambda \| R \|_F^2 =: U \\
\text{fix } U := V + R^*.
\]

- Use majorization-minimization (MM) (Hunter and Lange, 2000) to update

\[
H \gets H \cdot \frac{W^\top U}{W^\top WH} \quad \text{and} \quad W \gets W \cdot \frac{UH^\top}{WHH^\top}
\]

Initialization of \((W, H)\)

- Sample each entry independently from Half-Normal distribution (with variance parameter \(\gamma = 1\));
- Run 5 standard MM steps on \(V\) to obtain \(W_{\text{init}}\) and \(H_{\text{init}}\).

Termination (described in paper)
AT-NMF Algorithm (Update of \((W, H)\))

- After update of \(R\) to \(R^*\),
  
  \[
  \arg\min_{W,H \geq 0} \left\| V + R^* - WH \right\|_F^2 - \lambda \|R\|_F^2
  \]

  fix \(U := V + R^*\).

- Use majorization-minimization (MM) (Hunter and Lange, 2000) to update
  
  \[H \leftarrow H \cdot \frac{W^\top U}{W^\top WH}\]  
  \[W \leftarrow W \cdot \frac{UH^\top}{WHH^\top}\]

- Initialization of \((W, H)\)
  
  - Sample each entry independently from Half-Normal distribution (with variance parameter \(\gamma = 1\));

  - Run 5 standard MM steps on \(V\) to obtain \(W_{\text{init}}\) and \(H_{\text{init}}\).
AT-NMF Algorithm (Update of \((W, H)\))

- After update of \(R\) to \(R^*\),
  \[
  \arg\min_{W, H \geq 0} \| V + R^* - WH \|_F^2 - \lambda \| R \|_F^2 =: U
  \]
  fix \(U := V + R^*\).

- Use majorization-minimization (MM) (Hunter and Lange, 2000) to update
  \[
  H \leftarrow H \cdot \frac{W^T U}{W^T WH} \quad \text{and} \quad W \leftarrow W \cdot \frac{UH^T}{WHH^T}
  \]

- Initialization of \((W, H)\)
  - Sample each entry independently from Half-Normal distribution (with variance parameter \(\gamma = 1\));
  - Run 5 standard MM steps on \(V\) to obtain \(W_{\text{init}}\) and \(H_{\text{init}}\).

- Termination (described in paper)
Synthetic Dataset: Setup

- $\alpha \in \{0.1, 0.2, \cdots, 0.8, 0.9\}$ denotes the fraction of held-out entries.
- $\Gamma \subset \{1, \cdots, F\} \times \{1, \cdots, N\}$ is the set of held-out entries of $V$.
- $\hat{v}_{fn}$ is the prediction of $v_{fn}$. 

Synthetic Dataset: Setup

- $\alpha \in \{0.1, 0.2, \cdots, 0.8, 0.9\}$ denotes the fraction of held-out entries.
- $\Gamma \subset \{1, \cdots, F\} \times \{1, \cdots, N\}$ is the set of held-out entries of $V$.
- $\hat{v}_{fn}$ is the prediction of $v_{fn}$.
Our performance metric is the root mean-squared error (RMSE)

\[
\text{RMSE} := \sqrt{\frac{1}{|\Gamma|} \sum_{(f,n) \in \Gamma} (v_{fn} - \hat{v}_{fn})^2}
\]

Synthetic dataset of \( F = 100, N = 50, \) and \( K = 5. \)
Our performance metric is the root mean-squared error (RMSE)

\[
\text{RMSE} := \sqrt{\frac{1}{|\Gamma|} \sum_{(f,n) \in \Gamma} (v_{fn} - \hat{v}_{fn})^2}
\]

Synthetic dataset of \( F = 100, N = 50, \) and \( K = 5. \)

RMSE of Synthetic dataset

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>NMF</th>
<th>ANMF</th>
<th>AT-NMF (2)</th>
<th>AT-NMF (3)</th>
<th>AT-NMF (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>5.37 ± 0.02</td>
<td>6.78 ± 0.17</td>
<td>5.41 ± 0.12</td>
<td>5.11 ± 0.03</td>
<td>5.20 ± 0.02</td>
</tr>
<tr>
<td>0.4</td>
<td>5.62 ± 0.03</td>
<td>6.92 ± 0.17</td>
<td>5.54 ± 0.08</td>
<td>5.32 ± 0.09</td>
<td>5.42 ± 0.04</td>
</tr>
<tr>
<td>0.5</td>
<td>6.41 ± 0.01</td>
<td>7.44 ± 0.09</td>
<td>6.27 ± 0.11</td>
<td>6.05 ± 0.03</td>
<td>6.18 ± 0.02</td>
</tr>
<tr>
<td>0.6</td>
<td>6.74 ± 0.02</td>
<td>7.61 ± 0.09</td>
<td>6.47 ± 0.07</td>
<td>6.39 ± 0.03</td>
<td>6.53 ± 0.02</td>
</tr>
<tr>
<td>0.7</td>
<td>7.30 ± 0.01</td>
<td>7.99 ± 0.06</td>
<td>7.02 ± 0.04</td>
<td>6.94 ± 0.01</td>
<td>7.10 ± 0.02</td>
</tr>
<tr>
<td>0.8</td>
<td>7.87 ± 0.01</td>
<td>8.30 ± 0.06</td>
<td>7.69 ± 0.04</td>
<td>7.61 ± 0.03</td>
<td>7.71 ± 0.00</td>
</tr>
<tr>
<td>0.9</td>
<td>8.45 ± 0.01</td>
<td>8.58 ± 0.06</td>
<td>8.44 ± 0.02</td>
<td>8.34 ± 0.02</td>
<td>8.35 ± 0.02</td>
</tr>
</tbody>
</table>
$N = 2429$ facial images with $F = 361$ pixels.
CBCL Face Dataset: Parts Learned

- $N = 2429$ facial images with $F = 361$ pixels.
- Parts learnt when $\alpha = 0.1$
CBCL Face Dataset: Image Restoration

Image Restoration by AT-NMF

(a) Original Image $V$;
(b) Masked training image $V$;
(c) Adversary’s added-on masked image $R^*$;
(d) AT-masked image $V + R^*$ [Features (eyes, nose, lower cheeks) become more distinctive];
(e) Restored image using AT-NMF with $\lambda = 2$;
(f) Restored image using NMF.
Training losses when $\alpha = 0.5$
Hyperspectral Datasets

- It includes the Moffet (Jet Propulsion Lab) and Madonna (Sheeren et al., 2011) datasets with $F = 165$ and $F = 160$ respectively and $N = 2500$. 
Hyperspectral Datasets

- It includes the Moffet (Jet Propulsion Lab) and Madonna (Sheeren et al., 2011) datasets with $F = 165$ and $F = 160$ respectively and $N = 2500$.

- Effect of $\lambda$ on the RMSE
Conclusions and Future Work

- Formulation and algorithm for Adversarially-Trained NMF:

\[
\min_{W,H \geq 0} \max_{R \in \mathcal{R}} \| V + R - WH \|_F^2
\]

where \( \mathcal{R} = \{ R : \| R \|_F^2 \leq \epsilon, V + R \geq 0 \} \) or

\[
\min_{W,H \geq 0} \max_{R : V + R \geq 0} \| V + R - WH \|_F^2 - \lambda \| R \|_F^2.
\]
Conclusions and Future Work

- Formulation and algorithm for Adversarially-Trained NMF:
  \[
  \min_{W,H \geq 0} \max_{R \in \mathcal{R}} \| V + R - WH \|_F^2
  \]
  where \( \mathcal{R} = \{ R : \| R \|_F^2 \leq \epsilon, V + R \geq 0 \} \) or
  \[
  \min_{W,H \geq 0} \max_{R:V+R \geq 0} \| V + R - WH \|_F^2 - \lambda \| R \|_F^2.
  \]

- Other divergence measures beyond the Frobenius norm, e.g., \( \beta \)-divergence (Féotte et al., 2009).
Conclusions and Future Work

- Formulation and algorithm for Adversarially-Trained NMF:

\[
\min_{W,H \geq 0} \max_{R \in \mathcal{R}} \|V + R - WH\|_F^2
\]

where \(\mathcal{R} = \{R : \|R\|_F^2 \leq \epsilon, V + R \geq 0\}\) or

\[
\min_{W,H \geq 0} \max_{R:V+R \geq 0} \|V + R - WH\|_F^2 - \lambda \|R\|_F^2.
\]

- Other divergence measures beyond the Frobenius norm, e.g., \(\beta\)-divergence (Févotte et al., 2009).

- Different bounded sets \(\mathcal{R}\), e.g., \(\{R : \|R\|_{p,q} \leq \epsilon, V + R \geq 0\}\).
Conclusions and Future Work

- Formulation and algorithm for Adversarially-Trained NMF:

\[
\min_{W,H \geq 0} \max_{R \in \mathcal{R}} \| V + R - WH \|_F^2
\]

where \( \mathcal{R} = \{ R : \| R \|_F^2 \leq \epsilon, V + R \geq 0 \} \) or

\[
\min_{W,H \geq 0} \max_{R : V + R \geq 0} \| V + R - WH \|_F^2 - \lambda \| R \|_F^2.
\]

- Other divergence measures beyond the Frobenius norm, e.g., \( \beta \)-divergence (Févotte et al., 2009).

- Different bounded sets \( \mathcal{R} \), e.g., \( \{ R : \| R \|_{p,q} \leq \epsilon, V + R \geq 0 \} \).

- Online NMF (Lefèvre et al., 2011; Mairal, 2015)?


