Polytope Approximation and NMF

Moody T. Chu
(join work with Matthew M. Lin)

North Carolina State University

February 24, NISS Workshop

Outline

NMF Problem

- Given
  - A nonnegative matrix $Y \in \mathbb{R}^{m \times n}$,
  - A positive integer $p < \min\{m, n\}$,
- Find
  - Nonnegative matrices $U \in \mathbb{R}^{m \times p}$ and $V \in \mathbb{R}^{p \times n}$
  - Minimize the functional

$$f(U, V) := \frac{1}{2}\|Y - UV\|_F^2.$$

Basic Ideas

- Approximate a polytope by another polytope with fewer facets.
  - Reduce the number of vertices, but not the dimensionality.
- Work on the probability simplex.
  - Compact set with known boundary.
- Compute supporting hyperplanes in finitely many steps.
  - Find unique and global minimum per iteration.
- Applicable to NMF.
  - Might have applications to set estimation in pattern analysis, robot vision, and tomography — normally in $\mathbb{R}^3$. (Not in this talk)

Probability Simplex

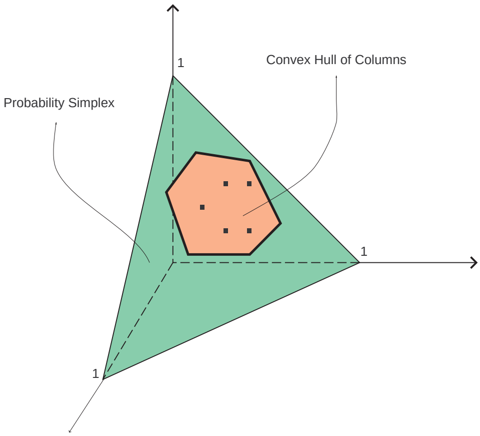
- Given $Y \in \mathbb{R}^{m\times n}$, define

$$
\begin{aligned}
\sigma(Y) &:= \operatorname{diag}\{\|\mathbf{y}_1\|_1, \ldots, \|\mathbf{y}_n\|_1\} \\
\vartheta(Y) &:= Y\sigma(Y)^{-1}.
\end{aligned}
$$

- Columns of $\vartheta(Y)$ are points on the probability simplex $\mathcal{D}_m$ in $\mathbb{R}^m$.

$$
\mathcal{D}_m := \left\{\mathbf{y} \in \mathbb{R}^m | \mathbf{y} \succeq 0, \mathbf{1}_m^\top \mathbf{y} = 1\right\},
$$

Convex hull of $\vartheta(Y) \in \mathbb{R}^{m\times n}$ with $m = 3$ and $n = 11$.



Minimal Convex Hull

- There is a smallest convex hull $\mathcal{C}$ containing all columns of $\vartheta(Y)$.

$$\begin{aligned} \mathcal{C} &:= \operatorname{conv}(\vartheta(Y)) = \operatorname{conv}(\vartheta(\widetilde{Y})), \\ \underbrace{\vartheta(Y)}_{m\times n} &= \underbrace{\vartheta(\widetilde{Y})}_{m\times p}\underbrace{Q}_{p\times n}. \end{aligned}$$

  - $\widetilde{Y} = [\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_p}]$ is a $m \times p$ submatrix of Y.
  - $Q \in \mathbb{R}^{p\times n}$ itself represents points in the simplex $\mathcal{D}_p$.
- This is an exact NMF of Y,

$$Y = \vartheta(Y)\sigma(Y) = \vartheta(\widetilde{Y})(Q\sigma(Y)).$$

  - $p \leq n$, but it might be that $p \geq m$.
  - Want $p << \min\{m, n\}$.

Converse

- If $Y = UV$ is an NMF of Y, then

$$Y = \vartheta(Y)\sigma(Y) = \vartheta(U)\vartheta(\sigma(U)V)\sigma(\sigma(U)V).$$

  - It must be such that

$$\begin{aligned} \vartheta(Y) &= \vartheta(U)\vartheta(\sigma(U)V), \\ \sigma(Y) &= \sigma(\sigma(U)V). \end{aligned}$$

  - WLOG, assume $\sigma(U) = I_n$, then

$$\begin{aligned} \vartheta(Y) &= \vartheta(U)\vartheta(V), \\ \sigma(Y) &= \sigma(V). \end{aligned}$$

Reformulation of NMF

- If $p < |\mathcal{C}|$, solving the NMF means minimizing

$$f(U,V) = \frac{1}{2}\|Y - UV\|_F^2 = \frac{1}{2}\left\| \left(\vartheta(Y) - U \underbrace{V\sigma(Y)^{-1}}_{W} \right) \sigma(Y) \right\|_F^2 .$$

  - Can consider W as the projection of the polytope $\vartheta(Y)$ onto the polytope $\mathrm{conv}(U)$ with respect to a weighted inner product.
  - Hahn-Banach theorem in a Hilbert space kicks in.
  - It is easier to work on the probability simplex.

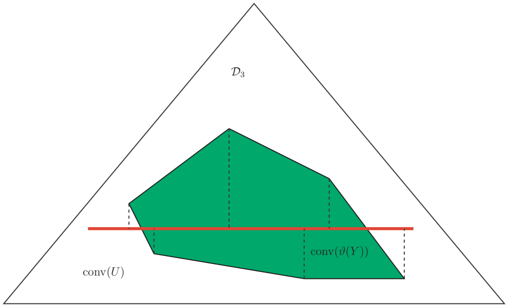
Convex Hull Fitting Problem

- Given $\vartheta(Y)$ and $p << \min\{m, n\}$,

$$
\begin{array}{ll}
\text{minimize} & g(U, W) = \frac{1}{2} \| \underbrace{\vartheta(Y)}_{m \times n} - \underbrace{U}_{m \times p} \underbrace{W}_{p \times n} \|_F^2, \\
\text{subject to} & U \in \partial \mathcal{D}_m, \quad W \succeq 0, \quad \mathbf{1}_p^\top W = \mathbf{1}_n^\top,
\end{array}
$$

  - $\partial \mathcal{D}_m$ stands for the boundary of $\mathcal{D}_m$.

Convex hull of $\vartheta(Y)$ and U in $\mathcal{D}_3$



Solving W

- For a fixed $U \in \mathbb{R}^{m \times p}$,

$$W = (U^\top U)^{-1}(U^\top \vartheta(Y) - \mathbf{1}_p \boldsymbol{\mu}^\top),$$

  - Lagrange multiplier,

$$\boldsymbol{\mu}^\top = \frac{\mathbf{1}_p^\top (U^\top U)^{-1} U^\top \vartheta(Y) - \mathbf{1}_n^\top}{\mathbf{1}_p^\top (U^\top U)^{-1} \mathbf{1}_p}.$$

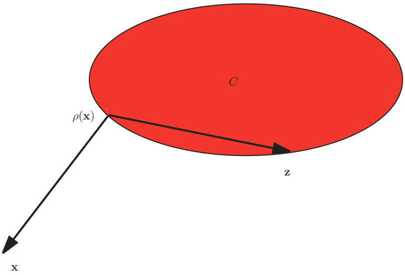
  - W may not be nonnegative.

Convex Coordinates

- Entries of W stands for the unique “coordinates" of $\vartheta(Y)$ in terms of U.
- The proximity map is guaranteed by the Hahn-Banach theorem.
  - Wolfe’s algorithm is available to find the nearest point of $\vartheta(\mathbf{y})$ on $\mathrm{conv}(U)$. (Wolfe’76)
  - More efficient recursive algorithm is also available. (Sekitani & Yamamoto’93)

Proximity Map to a Convex Set C

- Given $\mathbf{x}$, $\rho(\mathbf{x})$ = The nearest point on C to $\mathbf{x}$.
- Necessary and sufficient condition on $\rho(\mathbf{x})$:
  - $(\mathbf{x} - \rho(\mathbf{x}))^\top(\mathbf{z} - \rho(\mathbf{x})) \leq 0$ for all $\mathbf{z} \in C$.
  - $\|\rho(\mathbf{0})\|^2 \leq \rho(\mathbf{0})^\top \mathbf{z}$ for all $\mathbf{z} \in C$.



Hanh-Banach Theorem

- Two disjoint convex sets can be separated by a hyperplane.
- A hyperplane is determined by a normal vector **n** and a scalar c.

$$H(\mathbf{n}, c) := \{\mathbf{x} | \mathbf{n}^\top \mathbf{x} = c\}.$$

- A half space.

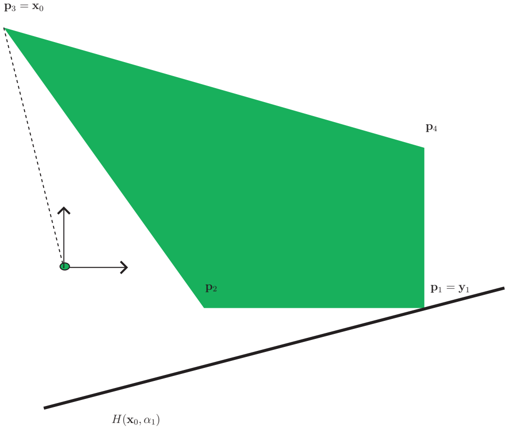
$$H^+(\mathbf{n}, c) := \{\mathbf{x} | \mathbf{n}^\top \mathbf{x} \geq c\}.$$

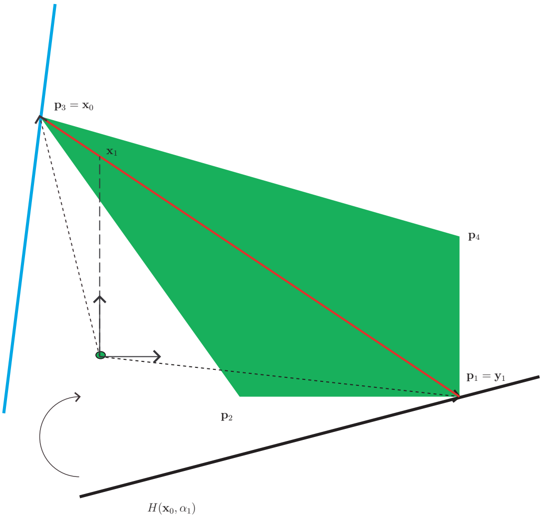
- Given C not containing the origin $\mathbf{0}$, $H(\rho(\mathbf{0}), \|\rho(\mathbf{0})\|^2)$ supports C in the sense that

$$C \subset H^+(\rho(\mathbf{0}), \|\rho(\mathbf{0})\|^2).$$

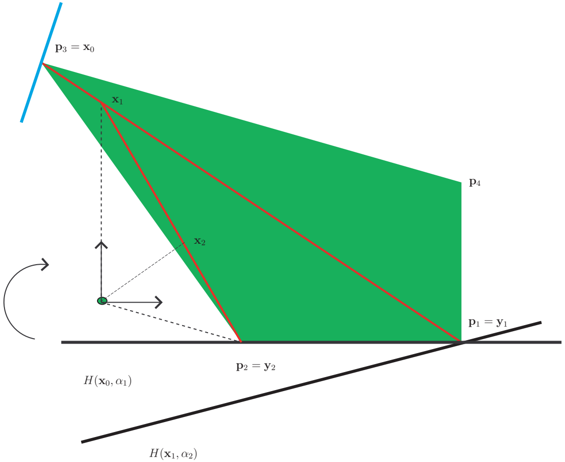
Sekitani and Yamamato Algorithm, $\widehat{\mathbf{x}} = \mathscr{N}(P)$

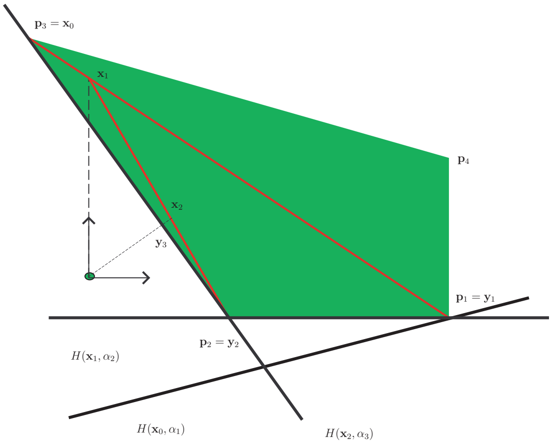
1. Start with $k := 1$ and an arbitrary point $\mathbf{x}_0$ from $\mathrm{conv}(P)$.
2. Find supporting hyperplane.
   - $\alpha_k := \min\left\{\mathbf{x}_{k-1}^\top \mathbf{p} | \mathbf{p} \in P\right\}$.
   - If $\|\mathbf{x}_{k-1}\|^2 \leq \alpha_k$, then $\widehat{\mathbf{x}} := \mathbf{x}_{k-1}$ and stop.
3. Recursion.
   - $P_k := \left\{\mathbf{p} | \mathbf{p} \in P \text{ and } \mathbf{x}_{k-1}^\top \mathbf{p} = \alpha_k\right\}$.
   - Call $\mathbf{y}_k := \mathscr{N}(P_k)$.
4. Check separation.
   - $\beta_k := \min\left\{\mathbf{y}_k^\top \mathbf{p} | \mathbf{p} \in P - P_k\right\}$.
   - If $\|y_k\|^2 \leq \beta_k$, then $\widehat{\mathbf{x}} := \mathbf{y}_k$ and stop.
5. Rotation.
   - $\lambda_k :=$ $\max\left\{\lambda | \left((1-\lambda)\mathbf{x}_{k-1} + \lambda \mathbf{y}_k\right)^\top \mathbf{y}_k \leq \left((1-\lambda)\mathbf{x}_{k-1} + \lambda \mathbf{y}_k\right)^\top \mathbf{p}, \mathbf{p} \in P - P_k\right\}$.
   - $\mathbf{x}_k := (1-\lambda_k)\mathbf{x}_{k-1} + \lambda_k \mathbf{y}_k$.
   - $k := k + 1$ and go to Step 2.


$\mathbf{p}_3 = \mathbf{x}_0$
$\mathbf{p}_4$
$\mathbf{p}_2$
$\mathbf{p}_1 = \mathbf{y}_1$
$H(\mathbf{x}_0, \alpha_1)$



$\mathbf{p}_3 = \mathbf{x}_0$
$\mathbf{x}_1$
$\mathbf{p}_4$
$\mathbf{p}_1 = \mathbf{y}_1$
$\mathbf{p}_2$
$H(\mathbf{x}_0, \alpha_1)$


$\mathbf{p}_3 = \mathbf{x}_0$
$\mathbf{x}_1$
$\mathbf{p}_4$
$\mathbf{x}_2$
$\mathbf{p}_1 = \mathbf{y}_1$
$\mathbf{p}_2 = \mathbf{y}_2$
$H(\mathbf{x}_0, \alpha_1)$
$H(\mathbf{x}_1, \alpha_2)$



$\mathbf{p}_3 = \mathbf{x}_0$
$\mathbf{x}_1$
$\mathbf{p}_4$
$\mathbf{x}_2$
$\mathbf{y}_3$
$\mathbf{p}_1 = \mathbf{y}_1$
$\mathbf{p}_2 = \mathbf{y}_2$
$H(\mathbf{x}_1, \alpha_2)$
$H(\mathbf{x}_0, \alpha_1)$
$H(\mathbf{x}_2, \alpha_3)$

Advantages

- Recursive in nature.
- Not based on simplicial decomposition.
  - No need to solve systems of linear equations.
- East to start.
  - Can start with an arbitrary point in $\mathrm{conv}(P)$.
  - Does not need an initial supporting hyperplane.
- Involves only matrix to vector multiplications.
- Find the unique global minimizer — the proximity map.
- Terminate in finite steps.
- Convex combination coefficients can be calculated.

Solving U

- Gradient is available.

$$\nabla_U g(U, W) := \frac{\partial g}{\partial U} = -(\vartheta(Y) - UW)W^\top.$$

- Assume $\mathbf{u}_i$ is on the jth facet, the projected gradient is easy to come by.

$$\nabla^j_{\mathbf{u}_i} g(U, W) := (I_m - A_j(A_j^\top A_j)^{-1}A_j^\top)\nabla_{\mathbf{u}_i} g(U, W). \tag{1}$$

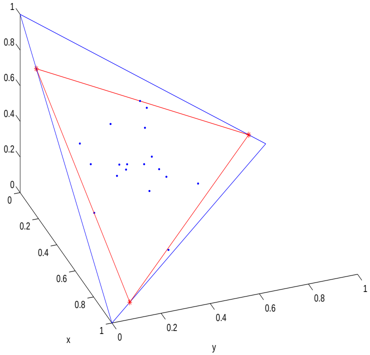
  - Projection matrix is easy to formulate.

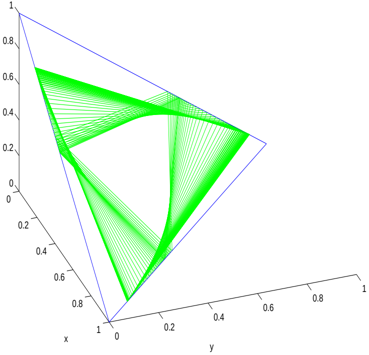
$$A_j(A_j^\top A_j)^{-1}A_j^\top = \frac{1}{m-1}\begin{bmatrix} 1 & \dots & 1 & 0 & 1 & \dots & 1 \\ \vdots & \ddots & & \vdots & & & \\ 1 & & 1 & 0 & 1 & & \\ 0 & & 0 & m-1 & 0 & \dots & 0 \\ 1 & & 1 & 0 & 1 & & 1 \\ \vdots & & & & & & \vdots \\ 1 & \dots & 1 & 0 & 1 & \dots & 1 \end{bmatrix},$$

Numerical Experiment

- Use the line search along the projected gradient direction to adjust U.
- U travels along the boundary of the simplex $\mathcal{D}_m$.
  - $\mathbf{u}_j$ may hit “ridges" of “vertices" of the simplex — can be detected.
  - Change facet is easy.
- Code is constructed and is under testing.

Triangle enclosing a prescribed set of points on $\mathcal{D}_3$





Nonnegative Matrix Factorization

- Similar approach can be generalized to NMF .
  - W is no longer on a simplex.
  - This becomes a weighted subspace approximation.
- The product UW should be interpreted as points of the *simplicial cone* of U.
  - By compactness, a truncated simplicial cone is enough.

NMF in $\mathbb{R}^2$

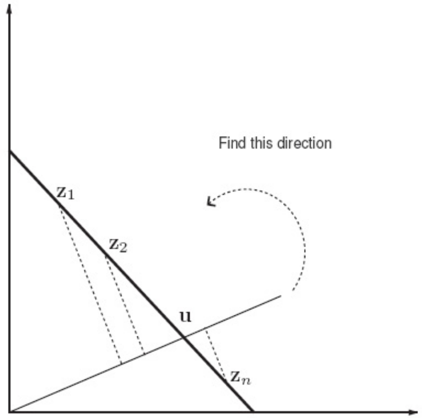
- Relationship between $\mathbf{u}$ and W in $\mathbb{R}^2$.

$$
\begin{aligned}
\mathbf{u} &= \sum_{i=1}^{n} \left(\frac{\sigma_i^2 w_i}{\sum_{i=1}^{n} \sigma_i^2 w_i^2} \vartheta(\mathbf{y}_i) - \frac{\sigma_i^2 w_i - \sigma_i^2 w_i^2}{2 \sum_{i=1}^{n} \sigma_i^2 w_i^2} \mathbf{1}_2 \right), \\
w_i &= \frac{\mathbf{u}^\top \vartheta(\mathbf{y}_i)}{\mathbf{u}^\top \mathbf{u}}, \quad i = 1, \ldots, n.
\end{aligned}
$$

- $\mathbf{u} w_i$ is precisely the projection of $\vartheta(\mathbf{y}_i)$ onto $\mathbf{u}$.
- w_i is guaranteed to be positive and is known as soon as $\mathbf{u}$ is given.
  - Not true in high-dimensional case.

Optimality in $\mathbb{R}^2$

- A geometric interpretation.



Nearest Point in Simplicial Cone

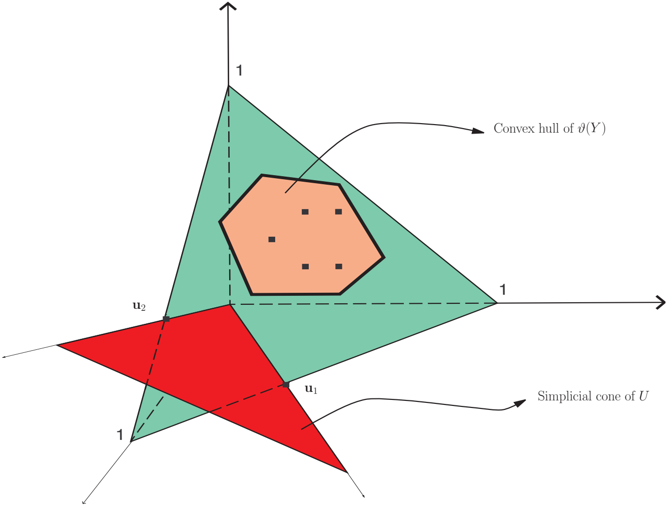
- Fix U, write

$$\begin{aligned} & f(U,V) = h(U,W) \\ := \quad & \frac{1}{2}\|(\vartheta(Y) - UW)\sigma(Y)\|_F^2 = \frac{1}{2}\sum_{i=1}^{n}\sigma_i^2\|\vartheta(\mathbf{y}_i) - U\mathbf{w}_i\|_2^2, \quad (2) \end{aligned}$$

  - If each term in (2) is minimized, the $h(U,W)$ is necessarily minimized.
  - Best approximate each column of $\vartheta(Y)$ within the simplicial cone of U.

Representing the Simplicial Cone

- With a large enough and fixed positive constant α, a truncated cone is given by
$$\widetilde{U} = [\mathbf{0}, \alpha\mathbf{u}_1, \ldots, \alpha\mathbf{u}_p].$$
- Columns of $\widetilde{U} \in \mathbb{R}^{m\times(p+1)}$ represent $p+1$ vertices of a polytope.
- Find the nearest point on $\mathrm{conv}(\widetilde{U})$ to $\vartheta(\mathbf{y}_i)$.
  - Can be done by Algorithm $\mathscr{N}$.
  - Obtain convex combination coefficients $\widetilde{W} \in \mathbb{R}^{(p+1)\times n}$ for $\vartheta(Y)$.


1
1
1
$\mathbf{u}_2$
$\mathbf{u}_1$
Convex hull of $\vartheta(Y)$
Simplicial cone of U

Unique Global Minimizer

- Decompose $\widetilde{W}$ into two blocks,

$$\widetilde{W} = \begin{bmatrix} \mathbf{w}_0^\top \\ W_0 \end{bmatrix}.$$

  - $\mathbf{w}_0^\top$ is the first row of $\widetilde{W}$.
  - $W_0 \in \mathbb{R}^{p \times n}$.
- No need of the origin.
  - Same points, but

$$\widetilde{U}\widetilde{W} = UW$$

  - $W = \alpha W_0$.
  - By construction, $W \succeq 0$.
  - W is no longer on $\mathcal{D}_p$.
- Given Y and U,

$$V = W\sigma(Y),$$

is the unique global minimizer to $f(U, V)$.

Updating U

- Can update U in exactly the same way as computing the optimal W.
- Consider

$$f(U,V) = \frac{1}{2}\|\left(\vartheta(Y^\top) - \vartheta(V^\top)\underbrace{(\sigma(V^\top)U^\top\sigma(Y^\top)^{-1})}_{\Phi}\right)\sigma(Y^\top)\|_F^2.$$

- Apply the procedures $\mathcal{N}$ to compute the unique and optimal simplicial combination coefficients $\Phi \in \mathbb{R}^{p\times m}$.
- The optimal U is given by

$$U = \left(\sigma(V^\top)^{-1}\Phi\sigma(Y^\top)\right)^\top.$$

Comparison with Lee and Seung

- Given U, compute V.
  - Chu and Lin algorithm,

$$V = W\sigma(Y),$$

  - Lee and Seung algorithm,

$$V^{+} = V. * (U^{\top} Y)./(U^{\top} UV),$$

- Given V, compute U.
  - Chu and Lin algorithm,

$$U = \left(\sigma(V^{\top})^{-1}\Phi\sigma(Y^{\top})\right)^{\top}.$$

  - Lee and Seung algorithm,

$$U^{+} := U. * (YV^{\top})./(UVV^{\top}),$$

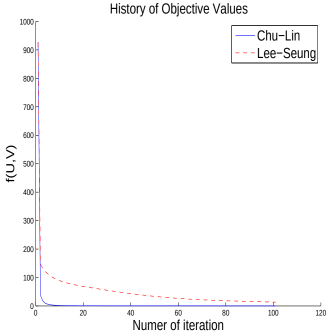
- Lee and Seung compute only the minimizer of an approximate and much simpler model.
- Chu and Lin compute the unique global minimizer.

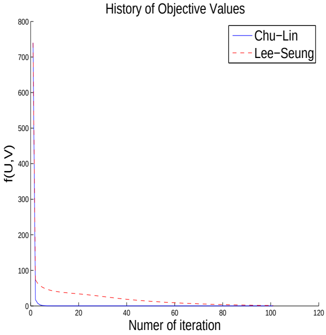
Numerical Experiment

- Test data.
  - Generate random nonnegative matrices $A \in \mathbb{R}^{m \times p}$ and $B \in \mathbb{R}^{p \times n}$ with $p < \min\{m, n\}$.
  - Let $Y = AB$ be the target data matrix.
- Can any NMF algorithm recover A and B from Y?

Accuracy

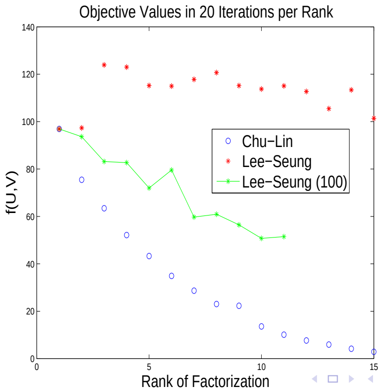
- Our method produces much closer approximation to Y, e.g., 3.3035×10^{-4} versus 1.1989, than Lee and Seung.





Improvement per Iteration

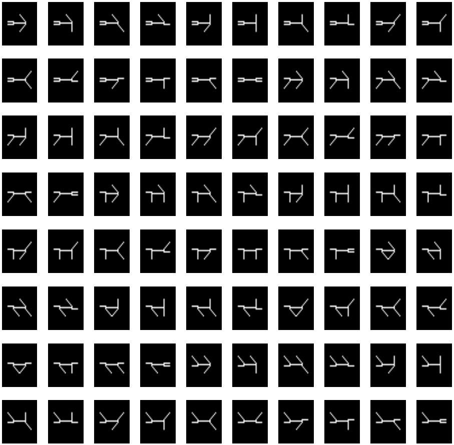
- Our method decreases the objective value more rapidly than Lee and Seung.



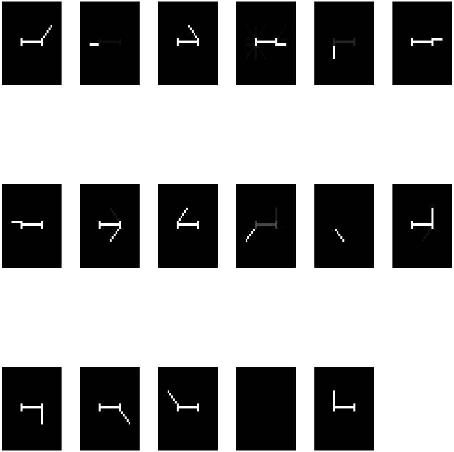
Swimmer Database

- A set of black-and-while stick figures satisfying the so called *separable factorial articulation criteria*.
- Each figure consists of a “torso" of 12 pixels in the center and four “limbs" of six pixels that can be in any one of four positions.
- With limbs in all possible positions, there are a total of 256 figures of dimension 32×32 pixels.
- Can the parts be recovered?

Eighty Swimmers



Seventeen Parts



Conclusion

- The notion of low dimensional polytope approximation is investigated in this talk.
  - The pull-back regulates the resulting polytopes to a more manageable compact set.
  - The proximity map can be calculated in finitely many steps.
- The proximity maps compute the unique global minimization in each alternating direction.
  - The best possible approximation per iteration.
  - Numerical experiments.
    - Smaller residual errors.
    - Fewer steps.

Future Work

- At present, the proximity map is accomplished column by column.
    - Less competitive in speed with the Lee-Seung algorithm which can be executed under BLAS3.
    - Possible to compute the proximity map for multiple columns simultaneously.
- A vectorization, if realizable, would be an added power to our method which in theory should produce the best possible approximation per alternating direction.